Régulation des transcrits, phénotypes intermédiaires entre génotype et phénotype : Focus sur les gènes à ARN longs non-codants (ARNlnc) chez la poule et les phénotypes liés aux lipides

Fabien Degalez

COLLEGE ECOLOGIE
DOCTORAL GEOSCIENCES
BRETAGNE AGRONOMIE ALIMENTATION

L'INSTITUT agro Rennes Angers

THÈSE DE DOCTORAT DE

L'INSTITUT AGRO RENNES ANGERS

ECOLE DOCTORALE N° 600
Écologie, Géosciences, Agronomie, Alimentation
Spécialité : *Génétique, génomique et bio-informatique*

Par
Fabien DEGALEZ

Régulation des transcrits, phénotypes intermédiaires entre génotype et phénotype

Focus sur les gènes à ARN longs non-codants (ARNlnc) chez la poule et les phénotypes liés aux lipides

Thèse présentée et soutenue à Rennes, le 13 décembre 2023 (IA, Amphi. Moule)
Unité de recherche : UMR 1348 PEGASE – INRAE / Institut Agro – Équipe Génétique et Génomique Avicole
Thèse N° : 2023-30 C-172

Rapporteurs avant soutenance :

Marie DE TAYRAC	PU-PH, CHU Pontchaillou, Rennes, France
Cédric NOTREDAME	Principal Investigator, CRG, Barcelone, Espagne

Composition du Jury :

Président :	Maria MANZANARES-DAULEUX	Professeure, Institut Agro Rennes-Angers
Examinateurs :	Cédric NOTREDAME	Principal Investigator, CRG, Barcelone, Espagne
	Marie DE TAYRAC	PU-PH, CHU Pontchaillou, Rennes, France
	Thomas DERRIEN	Chargé de recherche, CNRS, IGDR, Rennes, France
	Anamaria NECSULEA	Chargé de recherche, CNRS, GECO, Villeurbanne, France
Directrice de thèse :	Sandrine LAGARRIGUE	Professeure, Institut Agro Rennes-Angers

```
« ++++++++++
[>+>+++>+++++++>++++++++++<<<<-]
>>++++.>++++++.>---.<<--.>>++++++.
--.+++++++++.---------.
+++++++++++++++.-----------.
++++++++.++++.----------------.
<<++++++++++++.------------.
>>--.<<+++++++.>>++.+++++++++++++++.+
.<<-------.>>-----------------.
++++++++++++.--.+++.----.---.
++++++++.++++.----------------.<<.+.+. »
```

Telle une rengaine entêtante, les dires de ma directrice imprègnent mon jeune esprit.
Son credo a l'effet de sa forme sur ma pensée en gestation.
Façonnant avec patience mon âme scientifique.

Remerciements

J'aimerais, en premier lieu, sincèrement remercier ma directrice de thèse, Sandrine Lagarrigue, qui a, pendant ces quatre ans, assuré le bon déroulement des travaux tout en veillant à mon état d'esprit. Merci Sandrine. Merci d'avoir toujours cru en moi et de m'avoir accordé ta confiance. Cela a alimenté mon envie de donner le meilleur de moi-même et de toujours essayer d'avancer. Même si le découragement avait parfois raison de moi, tu as su me montrer les opportunités et me pousser à les saisir. Si des prises de bec ont pu quelques fois surgir, je pense qu'elles viennent avant tout de cette passion que nous partageons maintenant pour la science et que tu as réussi à me transmettre pendant ces années.

Je tiens également à remercier Frédéric Lecerf, Sophie Allais, Laetitia Lagoutte, Jean-Marc Fraslin, Pauline Philippe, Colette Désert, Bénédicte Lebez, Coralie Allain, Mathilde Doublet et Alexandre Hubert, membres de l'équipe présents sur le site de l'Agro et qui ont, à la fois par leur travail, mais aussi par leur bonne humeur, rendu cette thèse mémorable.

Frédéric, merci, grâce à toi, j'ai compris que, pour faire une thèse, il faut développer un sixième sens et être armé : toujours avoir des gants de boxe, un protège-dents et maintenir sa garde… sait-on jamais. Sophie, malgré quelques nuages annonçant occasionnellement une rapide tempête, tu restes avant tout un rayon de soleil réconfortant et une personne chaleureuse sur qui l'on peut compter. Laetitia, merci pour ta sincère attention et ta placidité qui cache tout de même bien souvent un petit diable farceur pouvant surgir à tout moment. Jean-Marc, merci pour la sagesse et la gentillesse dont tu as fait preuve et qui ne se perdront pas dans les tréfonds de ma mémoire tant cela m'a touché. Pauline, merci avant tout pour ton franc-parler et l'énergie que tu dégages, merci pour ton écoute, ton soutien infaillible à tout moment du jour et de la nuit et les différents exutoires que tu as su mettre en place pour garder le moral au plus haut. Colette, merci pour ta présence et nos échanges notamment durant le début de la thèse marquée par la période *covid*. Bénédicte, je cherche encore à comprendre comment tu arrives à être aussi efficace et concentrée malgré le vacarme que je suis capable de faire dans ton bureau, merci pour nos discussions et pour ta constante bonne humeur. Coralie, malgré ton arrivée plus récente, j'ai l'impression que tu as toujours été dans l'équipe, merci pour ton tact, ton énergie et ta pointe d'humour toujours présents. Mathilde,

merci d'avoir été mon compagnon de galère pendant une grande partie de ma thèse et d'avoir contribué activement à l'élaboration de nos plans les plus farfelus… même si je pense qu'il nous en reste encore un nombre non négligeable à mettre à exécution. Merci Alexandre pour tes multiples anecdotes et récits qui ont su animer les repas et les pauses et que j'écoute bien souvent avec grand plaisir. Je te souhaite, ainsi qu'à Mathilde, de réussir vos thèses comme vous l'entendez. Pour finir, j'aimerais adresser une pensée émue à destination de Morgane Boutin.

Mes remerciements vont également à Frédéric Jehl, Kévin Muret, Florian Herry et Yuna Blum, anciens doctorants du laboratoire avec qui j'ai eu l'opportunité de travailler et discuter dans un cadre tout autant professionnel que personnel.

Je tiens plus particulièrement à remercier Frédéric Jehl qui, alors qu'il terminait sa thèse, a pris en charge mon encadrement en stage, puis m'a grandement assisté en début de thèse. Merci Frédéric, ta persévérance, ton attention à chaque détail, mais aussi ta folie, m'ont poussé à suivre assez trivialement ton exemple durant ces quatre années.

J'aimerais remercier Christian Diot, Frédéric Hérault, Romain Philippe, Nicolas Bédère et Lorry Bécot, membres de l'équipe localisés sur le site de Saint-Gilles, qui ont su apporter conseils et bienveillance tout au long de la thèse.

Mes remerciements s'adressent aussi aux collègues des autres unités et départements, et plus particulièrement à Clara Lambard, Anne-lise Jacquot, Lucile Montagne, Yannick Le Cozler, Marie-Emmanuelle Blanchard, Jocelyne Flament, Sophie Brajon, Pierre-Guy Marnet, Sylvie Fortin, Sylvie Bonnassie, Jacques Portanguen, Tino Jamme et Auxane Hammon, présents au sein du bâtiment et avec qui j'ai pu entreprendre de nombreuses discussions aux sujets très hétéroclites et passant bien souvent en un éclair du coq à l'âne.

I would like to thank Professor Stephen Montgomery who welcomed me into his laboratory at Stanford University for a three-month exchange, but also Pagé Goddard, Rachel Ungar, Emily Greenwald and Tanner Jensen, who provided me real support on both a professional and personal level. I am particularly grateful to Emily and Tanner, whose kindness, sincerity and sensitivity have made them true friends. I would also like to thank all the people I met during this time.

Je tiens à remercier Sylvain Foissac, Mathieu Emily, Thomas Derrien, Frédérique Pitel, Sophie Allais et Frédéric Lecerf qui ont accepté d'être membres de mon comité de thèse. Merci pour votre suivi et pour les parfois longues, mais toujours intéressantes, discussions qui en ont résulté.

J'adresse mes sincères remerciements à Marie De Tayrac, Cédric Notredame, Maria Manzanares-Dauleux, Anamaria Necsulea et Thomas Derrien qui ont accepté d'être membres du jury de la présente thèse. Merci d'avoir relu et corrigé avec attention celle-ci, et ce, dans des délais assez courts. Merci également pour les discussions ayant suivi la soutenance qui ont été très constructives et instructives.

Je tiens enfin à remercier les financeurs de cette thèse : la Région Bretagne et le département de Génétique Animale de l'INRAE.

Plus personnellement, je souhaite remercier les membres de ma famille qui ont, parfois, tant bien que mal, tenté de comprendre les travaux que je menais dans le cadre de cette thèse et les implications tant professionnelles que personnelles d'un tel projet. Je tiens à vous remercier pour votre soutien dans les différentes étapes, du lancement de celle-ci au pot la clôturant en passant par votre présence à la soutenance.

J’aimerais remercier les personnes qui ont dû supporter à mes côtés et au quotidien cette thèse au prix parfois de certains sacrifices et plusieurs crises d’humeur. Merci Hélène de m’avoir accompagné durant le début de la thèse et d’avoir fait preuve de compréhension vis-à-vis du rythme que je m’imposais. Merci également de m’avoir soutenu jusqu’aux derniers moments de celle-ci. Manon, merci d’avoir été un pilier inébranlable et de m’avoir soutenu entièrement dans ce projet. Merci de m’avoir apporté douceur, optimisme et affection, et cela, même dans les moments les plus durs et les plus cruciaux.

Je ne pourrais oublier mes amis qui ont fait partie intégrante de cette thèse et qui m’ont permis de la mener correctement jusqu’aux derniers moments. Ma première pensée est bien évidemment adressée aux « Dindes Swag », le fidèle groupe des copains de prépa et maintenant famille infaillible répondant fièrement présent dans toutes les situations. Merci également aux « Copains D’abord », qui, en plus de m’avoir offert des temps de pause autour de quelques moments festifs, m’ont accompagné depuis le premier jour à ACO pour forger des souvenirs d’école inoubliables. Au risque d’en oublier, je tiens à remercier plus personnellement Nathalie, Romain, Cynthia, Paul, Roxane, Adrien, Vincent, Antoine, Laurine, Fanny, Estelle, qui m’ont très largement épaulé, notamment durant cette période. Malgré une distance qui s’est installée, mes pensées vont également à mes amis d’enfance et plus particulièrement à Julien qui a toujours cru en moi en me donnant tout son soutien et m’a poussé à donner le meilleur de moi-même.

Pour finir, j’aimerais clôturer ces remerciements en m’adressant à l’ensemble des personnes que j’ai eu le plaisir de côtoyer et avec qui j’ai pu partager quelques moments de vie.

Valorisations des travaux de thèse

À la date de soumission du manuscrit, les valorisations réalisées au cours de la présente thèse, présentées par ordre chronologique, sont les suivantes :

Publications acceptées :

- Jehl F*, **Degalez F***, Bernard M*, Lecerf F, Lagoutte L, Désert C, Coulée M, Bouchez O, Leroux S, Abasht B, Tixier-Boichard M, Bed'hom B, Burlot T, Gourichon D, Bardou P, Acloque H, Foissac S, Djebali S, Giuffra E, Zerjal T, Pitel F, Klopp C and Lagarrigue S (2021). RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Frontiers in Genetics. doi: 10.3389/fgene.2021.655707

- **Degalez F***, Jehl F*, Muret K, Bernard M, Lecerf F, Lagoutte L, Désert C, Pitel F, Klopp C and Lagarrigue S (2021). Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained. Frontiers in Genetics. doi: 10.3389/fgene.2021.659287

- Lagarrigue S, Lorthois M, **Degalez F**, Gilot D, Derrien T (2022). LncRNAs in domesticated animals: from dog to livestock species. Mammalian Genome. doi: 10.1007/s00335-021-09928-7

- **Degalez F**, Muret K, Lagarrigue S (2023). Evolution of protein coding and long non coding genes of the chicken genome through the different genome assemblies and their associated annotations. Cytogenetic and Genome Research. doi: 10.1159/000529376 – Inscrit dans le cadre du "Fourth Report on Chicken Genes and Chromosomes 2022" mené par Jacqueline Smith.

Publications en révision :

- **Degalez F**, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F, Acloque H, Giuffra E, Pitel F, Lagarrigue S. (2023). Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues. bioRxiv. doi: 10.1101/2023.08.18.553750 – *Soumis à Scientific Reports*

- Guan D, …, **Degalez F**, Lagarrigue S, …, Zhou H, Fang L. – The ChickenGTEx Consortium. (2023). The ChickenGTEx pilot analysis: a reference of regulatory variants across 28 chicken tissues. doi: 10.1101/2023.06.27.546670 – *Soumis à Nature Genetics*

Publications en finalisation d'écriture :

- Ungar R*, Goddard P*, Jensen T, **Degalez F**, Smith K, Jin C, Bonner D, Bernstein J, Wheeler M, Montgomery S. (2023). Impact of genome build on RNA-seq interpretation and rare disease diagnosis – *Journal visé : Nature Genetics*

- **Degalez F**, Bardou P, Lagarrigue S (2023). GEGA (Gallus Enriched Gene Annotation): an online tool gathering genomics and functional information across 47 tissues for 78,323 protein-coding genes and lncRNAs including Ensembl & Refseq genome annotation – *Journal visé : Nucleic Acid Research*

(e)-Posters :

- Jehl F*, **Degalez F***, Bernard M, Lecerf F, Coulee M, Zerjal T, Pitel F, Klopp C, Lagarrigue S. (2020). Genomic SNP detection by RNA-seq: lessons from multi-tissue & multi-population data analysis in chickens. ePoster présenté aux "Open Day of Computational Biology and Mathematics" (JOBIM), Montpellier, France.

- **Degalez F**, Lagoutte L, Lecerf F, Vlach M, Lagarrigue S. (2022). Gene orthology detection for long non-coding RNA. Poster présenté aux "Open Day of Computational Biology and Mathematics" (JOBIM), Rennes, France.

- Greenwald E, Park J, **Degalez F**, Jain N, Zheludv V, Artiles J, Jeong D, Wahba L, Rodrigues K, Galls D, Yin W, Fire A. (2022). RNA monsters generated by mitochondrial RNA polymerase. Poster présenté lors de l' "American Society of Human Genetics" (ASHG), Los Angeles, California, United-States.

- Goddard P, Ungar R, Jensen T, Marwaha S, Bonner D, **Degalez F**, Smith K, Montgomery S. (2022). Genome reference impacts RNA-seq interpretation and rare disease diagnosis. Poster présenté lors de l' "American Society of Human Genetics" (ASHG), Los Angeles, California, United-States.

- **Degalez F**, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F, Giuffra E, Zerjal T, Pitel F, Lagarrigue S. (2023). A lncRNA gene-enriched atlas for GRCg7b chicken genome and its functional annotation across 47 tissues. Poster présenté dans la session spécialisée "Avian Genetics and Genomics" au 39ème congrès de l' "International Society for Animal Genetics" (ISAG), Cape Town, South Africa.

- **Degalez F**, Allain C, Lagoutte L, Lagarrigue S. (2023). Gene orthology detection for long noncoding RNA (lncRNA). Poster présenté dans la session spécialisée "Animal Epigenetics" au 39ème congrès de l' "International Society for Animal Genetics" (ISAG), Cape Town, South Africa.

- **Degalez F**, Bardou P, Lagoutte L, Allain C, Lagarrigue S. (2023). LncRNA analysis in response to diet changes in chicken liver. Poster présenté dans la session spécialisée "Avian Genetics and Genomics" au 39ème congrès de l' "International Society for Animal Genetics" (ISAG), Cape Town, South Africa.

- **Degalez F**, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F, Giuffra E, Zerjal T, Pitel F, Lagarrigue S. (Sept 2023). A lncRNA gene-enriched atlas for GRCg7b chicken genome using Ensembl, RefSeq and two FAANG database. Poster présenté dans la session spécialisée "EuroFAANG: genotype-to-phenotype research across Europe and beyond" pour le "74th European Federation of Animal Science Meeting" (EAAP), Lyon, France.

Communications orales :

- Jehl F*, **Degalez F***, Bernard M, Lecerf F, Coulee M, Zerjal T, Pitel F, Klopp C, Lagarrigue S. (2020). Genomic SNP detection by RNA-seq: lessons from multi-tissue & multi-population data analysis in chickens. Exposé oral lors des "Open Day of Computational Biology and Mathematics" (JOBIM), Montpellier, France.

- Jehl F, **Degalez F**, Bernard M, …, Klopp C, Lagarrigue S. (2022). RNA-seq data for detecting reliable SNPs & genotypes in livestock species: interest for coding variant characterization and cis-regulation analysis by allele-specific expression. Exposé oral dans la session spécialisée "Molecular Genetics" au "World's Poultry Congress" (WPC), Paris, France.

- Goddard P, Ungar R, Jensen T, Marwaha S, Bonner D, **Degalez F**, Smith K, Montgomery S. (2022). Genome reference impacts RNA-seq interpretation and rare disease diagnosis. Exposé oral lors de l'"American Society of Human Genetics" (ASHG), Los Angeles, California, United-States.

- **Degalez F**, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F, Giuffra E, Zerjal T, Pitel F, Lagarrigue S. (2023). A lncRNA gene-enriched atlas for GRCg7b chicken genome and its functional annotation across 47 tissues. Exposé oral dans la session spécialisée "Avian Genetics and Genomics" au 39ème congrès de l' "International Society for Animal Genetics" (ISAG), Cape Town, South Africa.

Liste des abréviations

ASE : *Allele Specific Expression* ; Expression allèle-spécifique
bp : *base pair* ; Paire de bases
CAGE : *Cap Analysis of Gene* Expression ; Analyse de l'expression des gènes en 5'
CattleGTEx : *Cattle Genotype-Tissue Expression*
CGNC : *Chicken Gene Nomenclature Committee* ; Comité de nomenclature des gènes de la poule
ChickenGTEx : *Chicken Genotype-Tissue Expression*
CPU : *Central Processing Unit* ; Unité centrale de calcul
CR : *Call-Rate* ; Taux d'appel
DEG : *Differentially Expressed Gene* ; Gène différentiellement exprimé
DNAseq : *DNA-sequencing* ; Séquençage de l'ADN
DP : *Depth* ; Profondeur
EBI : *European Bioinformatics Institute* ; Institut européen de bio-informatique
EMBL : *European Molecular Biology Laboratory* ; Laboratoire européen de biologie moléculaire
Ensembl : *EMBL-EBI Ensembl/GENCODE*
eQTL : *expression Quantitative Trait Loci*
FAANG : *Functional Annotation of Animal Genomes* ; Annotation fonctionnelle des génomes animaux
FarmGTEx : *Farm Genotype-Tissue Expression*
FDR : *False Discovery Rate* ; Taux de fausses découvertes
FPKM : *Fragments Per Kilobase per Million mapped reads*
GEGA : *Gallus Enriched Gene Annotation* ; Annotation génique enrichie pour la poule
GT : *Genotype* ; Génotype
GPU : *Graphical Processing Unit* ; Unité de traitement graphique
GTEx : *Genotype-Tissue Expression*
GWAS : *Genome-Wide Association Study* ; Étude d'association pangénomique
HGNC : *HUGO Gene Nomenclature Committee* ; Comité de nomenclature des gènes HUGO
HUGO : *HUman Genome Organisation* ; Organisation du Génome Humain
INDEL : *INsertion DELetion*
LD : *Linkage Desequilibrium* ; Déséquilibre de liaison
lncRNA : *Long Non Coding RNA* ; ARN long non-codant
MAF : *Minor Allele Frequency* ; Fréquence de l'allèle mineur
MANE : *Matched Annotation from NCBI and EMBL-EBI* ; Annotation conjointe du NCBI et de l'EMBL-EBI
MGNC : *Mouse Gene Nomenclature Committee* ; Comité de nomenclature des gènes de la souris
MNV : *Multi-Nucleotide Variants* ; Variants multi-nucléotidiques
molQTL : *molecular Quantitative Trait Loci*

NCBI : *National Center for Biotechnology Information* ; Centre américain pour les informations biotechnologiques
NGS : *Next Generation Sequencing* ; Séquençage de nouvelle génération
NIH : *National Institutes of Health* ; Instituts américains de la santé
OMIA : *Online Mendelian Inheritance in Animals*
ORF : *Open Reading Frame* ; Cadre de lecture ouvert
PCG : *Protein Coding Gen*e ; Gène codant des protéines
PigGTEx : *Pig Genotype-Tissue Expression*
QTL : *Quantitative Trait Loci* ; Locus de caractères quantitatifs
RefSeq : *NCBI-RefSeq* ; *NCBI Reference Sequence Database*
RLE : *Relative Log Expression*
RGNC : *Rat Genome Nomenclature Committee* ; Comité de nomenclature des gènes du rat
RNAseq : *RNA-sequencing* ; Séquençage de l'ARN
RPKM : *Reads Per Kilobase per Million mapped reads*
sb-eQTL : *sex-biased-eQTL*
SNP : *Single Nucleotide Polymorphism* ; Polymorphisme Nucléotidique
T2T : *Telomere To Telomere*
TAD : *Topological Associated Domain* ; Domaine topologique associé
TF-eQTL : *Transcription-Factor-eQTL*
TMM : *Trimmed Mean of M-values*
TPM : *Transcript Per Million*
TS : *Tissu Specific* ; Spécifique de tissus
TSS : *Transcription Start site* ; Site d'initiation de la transcription
TUCR : *Transcribed UltraConserved Regions* ; Régions transcrites ultraconservées
TWAS : *Transcriptome-Wide Association Study* ; Étude d'association à l'échelle du transcriptome
UTR : *UnTranslated Region* ; Région non traduite
VEP : *Variant Effect Predictor*
VGNC : *Vertebrate Gene Nomenclature Committee* ; Comité de nomenclature des gènes de vertébré
ZNC : *Zebrafish Nomenclature Committee* ; Comité de nomenclature des gènes du poisson zèbre

Table des matières

Objectifs de la thèse

La thèse présentait trois objectifs, **le premier** consistait en l'amélioration de l'annotation des gènes du génome de la poule, notamment pour les gènes régulateurs à ARN longs non-codants (lncRNA). **Le second** était d'identifier et de caractériser les variants à l'échelle du génome et plus particulièrement par séquençage ARN. Enfin, **le troisième** et dernier objectif était de s'appuyer sur les résultats des deux précédents afin d'étudier par des études d'association pangénomique (GWAS pour *Genome-Wide Association Studies*) la composante génétique des caractères complexes depuis l'expression des gènes (focus sur le foie) jusqu'à des phénotypes macroscopiques (focus sur le gras).

Plus précisément, le **premier objectif** a été de contribuer à l'annotation du génome de la poule Pour ce faire, nous avons conçu une annotation du génome enrichie en lncRNA en considérant la dernière version de l'assemblage du génome GRCg7b considérée comme référence depuis 2022 par « EMBL-EBI Ensembl/GENCODE » (abrégé en Ensembl). Cette annotation du génome de la poule incluant, entre autres, les bases de données de référence Ensembl et « NCBI-RefSeq » (abrégé en RefSeq) est également accompagnée d'une annotation fonctionnelle des gènes du génome au travers de profils d'expressions établis à partir de 1400 échantillons et 47 tissus représentant globalement les divers systèmes physiologiques (Résultats §1.1). Afin de permettre un accès aisé à cette ressource, un outil en ligne, GEGA (pour *Gallus Enriched Gene Annotation* – Annotation génique enrichie pour la poule) a été développé, permettant de filtrer les gènes selon des critères souhaités, mais aussi d'observer les profils d'expressions à plusieurs niveaux : que ce soit à l'échelle de l'ensemble des tissus (inter-tissus), à l'intérieur même d'un tissu afin d'observer la variabilité individuelle (intra-tissus) ou encore selon certaines conditions telles que l'âge ou le sexe (Résultats §1.2). Durant de ma mobilité dans le laboratoire du Pr. Stephen Montgomery, des travaux complémentaires sur l'impact du choix des assemblages humains et de leurs annotations associées sur la détection d'expressions aberrantes pour l'étude des maladies rares ont également été réalisés (Résultats §1.3). Enfin, toujours dans l'objectif d'améliorer l'annotation des gènes, un *pipeline* permettant de s'intéresser à la conservation multi-espèces des lncRNA a été développé. Il a été appliqué à un panel de 13 espèces incluant le poisson zèbre, des espèces aviaires dont la poule, les mammifères domestiqués, la souris et l'humain et a permis

de mettre en évidence de potentiels liens d'orthologie pour plusieurs milliers de lncRNA (Résultats §1.4). Notons que ces derniers travaux sur l'orthologie des lncRNA ont été initiés durant le dernier trimestre 2021, sur le génome de référence GRCg6a. Les analyses sont donc à réitérer sur le génome de référence GRCg7b.

Le **deuxième objectif** a été d'identifier et de caractériser des polymorphismes nucléotidiques (SNP, pour *Single Nucleotide Polymorphism*) pouvant servir à des études GWAS menées, par la suite, dans le but d'identifier les gènes et/ou variants expliquant une partie de la variation des caractères observés. Ainsi, nous nous sommes tout d'abord intéressés à la détection de tels SNP à partir de données issues de séquençage ARN (RNAseq, pour *RNA sequencing*) servant également à quantifier l'expression des gènes. Ce travail a consisté à la mise en place d'un *pipeline* et de filtres afin d'extraire des SNP (9,5 M) et leurs génotypes associés considérés comme fiables, ceci en comparaison avec ceux obtenus par séquençage ADN 20X (DNAseq 20X, pour *DNA sequencing*) pour les mêmes individus. Les génotypes ainsi extraits ont ensuite été utilisés pour évaluer l'intérêt du RNAseq pour *i)* des analyses d'expression allèle-spécifique (ASE, pour *Allele Specific Expression*), *ii)* l'exploration de la diversité génétique entre populations, et, *iii)* la prédiction des impacts fonctionnels des SNP sur les protéines (Résultats §2.1). Ces analyses ont été menées sur 744 RNAseq répartis au sein de 11 populations variées de poules. Considérant la prédiction des impacts fonctionnels, nous avons pris en considération, dans une seconde étude, les informations de phase entre SNP d'un même codon, permettant ainsi d'affiner les prédictions faites usuellement SNP par SNP. Un *pipeline* permettant de recalculer les conséquences pour ce type de cas a alors été développé et testé sur les 9,5 millions de SNP détectés (Résultats §2.2). Pour finir, dans le cadre du projet européen GEroNIMO et l'ARN EFFICACE, une population commerciale de poules pondeuses a été génotypée en routine par des puces 600K et 60K. Les marqueurs de ces puces étant selon l'assemblage galgal5, les premiers travaux ont consisté en la conversion selon le dernier assemblage GRCg7b puis, par la suite, en un contrôle qualité permettant d'extraire les génotypes fiables. Pour les individus génotypés par la puce 60K, une imputation vers la puce 600k a également été réalisé afin de densifier le nombre de marqueurs disponibles (Résultats §2.3).

Le **troisième objectif** enfin a été de combiner les données d'annotation du génome de la poule, obtenues lors du **premier objectif,** avec les SNP détectés dans le cadre du **second objectif** dans le but de localiser par GWAS les régions du génome impliquées dans la variabilité

phénotypique que ce soit pour des phénotypes de type « expression des gènes » ou encore des phénotypes macroscopiques plus complexes liés au métabolisme des lipides. Ces régions responsables d'une part de la variation d'un phénotype sont appelées eQTL (*expression Quantitative Trait Loci)* et QTL (*Quantitative Trait Loci)* pour les caractères expressionnels et macroscopiques respectivement. Nous avons ainsi participé au projet international pilote ChickenGTEx (*chicken Genotype-Tissue Expression*), lui-même intégré dans le projet FarmGTEx, visant à identifier des régions régulatrices et régulées du génome de la poule au travers de 28 tissus provenant de populations hétérogènes (Résultats §3.1). De manière similaire, pour la population commerciale de poules pondeuses évoquée dans le deuxième objectif, nous nous sommes intéressés à la détection de régions eQTL en questionnant l'impact des co-variables ainsi que l'effet d'échantillonnage sur la puissance des résultats obtenus. Ces eQTL ont été mis en parallèle de QTL liés à des phénotypes en lien avec le métabolisme des lipides afin de favoriser l'identification de gènes/variants candidats causaux dans le foie (Résultats §3.2).

Afin de remettre dans leurs contextes ces objectifs, l'introduction qui suit est composée de trois parties. Dans la **première partie**, nous exposons différents aspects de l'annotation des gènes en commençant par une brève histoire du concept de gène permettant de comprendre les contours de la définition actuelle ainsi que les apports des nouvelles technologies de séquençage (NGS pour *Next Generation Sequencing*) et notamment l'arrivée à bas coût du RNAseq ayant permis la considération des lncRNA (Introduction §1.1). Ainsi, les méthodes permettant leur prédiction, leurs modes d'actions et les différentes classifications proposées dans la littérature font l'objet d'un second chapitre qui s'appuie sur deux revues publiées dans le cadre de la thèse (Introduction §1.2). Pour finir, les concepts en lien avec la conservation des gènes entre les différentes espèces sont présentés et les spécificités propres aux lncRNA et aux gènes codants des protéines (PCG pour *Protein Coding Gene*) sont commentées (Introduction §1.3). Dans la **seconde partie**, nous présentons la méthode du RNAseq et dressons un portrait des différents usages possibles (Introduction §2.1). Par la suite, nous nous focalisons davantage sur la quantification de l'expression des gènes via cette méthode notamment en présentant les différentes métriques pouvant être employées et sans oublier celles permettant d'évaluer la tissue-spécificité (Introduction §2.2). Dans la **troisième et dernière partie**, nous nous intéressons aux variants génétiques et à leur utilisation en GWAS

pour élucider la relation complexe entre génotype et phénotype. Après une remise en perspective sur l'impact de la génétique sur les caractères observables, nous proposons une synthèse des caractéristiques majeures des variants identifiés par séquençage haut débit et impliquées dans les caractères d'intérêts (Introduction §3.1). La majorité des SNP détectés par GWAS pour des caractères complexes étant localisés dans des séquences régulatrices agissant sur l'expression de gènes, nous verrons comment la considération de phénotypes expressionnels peut faciliter l'identification de zones du génome (voir de gènes) responsables de caractères plus complexes (Introduction §3.2). Pour finir, nous présentons les méthodes dites de colocalisation permettant de mettre en parallèle les eQTL et QTL obtenus (Introduction §3.3).

À la suite de cette « Introduction », les travaux en lien avec nos objectifs sont présentés dans la partie « Résultats ». Pour la majorité des paragraphes en lien avec un article publié, soumis ou rédigé, une synthèse en français de l'article est proposée. Notons que ce résumé s'appuie sur les figures de l'article associé nécessitant donc de s'y référer selon les indications fournies. Pour les travaux sans valorisation par le biais d'un article, les résultats produits sont exposés sous la forme d'un paragraphe davantage détaillé. Dans l'ensemble des cas, les numéros de référence correspondent à ceux utilisés dans le cadre du présent manuscrit et non à ceux des articles publiés.

Pour finir, afin de mieux les comprendre, d'évaluer leur importance, mais aussi pour les remettre dans un contexte global, ces résultats sont discutés dans la partie « Discussions et perspectives » selon deux paragraphes. Le premier se focalise sur les efforts collectifs nécessaires pour améliorer les annotations des modèles géniques et leurs annotations fonctionnelles, notamment pour les lncRNA. Le second, pour finir, s'attarde sur les interprétations liées aux analyses conjointes des QTL et eQTL.

L'ensemble des références utilisées dans le manuscrit est rendu disponible dans la partie « Bibliographie » dédiée.

Introduction

1. Annotation des gènes du génome, le RNAseq haut-débit un tournant

1.1. L'évolution du concept de gènes, de l'expérience de pensée antique à la génomique de masse

1.1.1. Approche conceptuelle de l'hérédité à l'Antiquité : des humeurs d'Hippocrate à la forme et la matière d'Aristote

Avant de traiter de la notion de gène, il convient de s'attarder auparavant sur le concept d'hérédité ayant servi de base à la conceptualisation et à l'évolution de sa définition au cours de l'Histoire.

L'hérédité, du latin *hereditas* (« ce dont on hérite, succession ») [1], se définit comme la transmission de caractères et traits physiques des parents vers leur descendance. Si ce concept présente à première vue une certaine dualité antinomique impliquant la constance d'une espèce d'une génération à une autre d'une part, mais la variation entre les individus d'une même espèce d'autre part, ces deux éléments se révèlent en réalité complémentaires. Les premières évocations de ce phénomène datent de l'Antiquité, même si les mécanismes sous-jacents n'étaient pas encore établis. Les philosophes et scientifiques tentent alors d'expliquer les similitudes et les différences entre les descendants et les parents. C'est ainsi qu'Hippocrate de Kos (460-377 av. J.-C.), considéré de nos jours comme le « père de la médecine », applique le concept des « humeurs » au domaine de la médecine en lui donnant une notion héréditaire [2]. Ce concept d'humeurs définit comme des entités chimiques régulant le comportement, existait déjà auparavant, mais a été reconsidéré par Hippocrate dans son traité « *The Nature of Man* » [3–5] où il en définit quatre, associés aux quatre éléments : *i)* la bile noire associée à la mélancolie et à la terre, *ii)* la bile jaune associé à une nature colérique, ambitieuse ou encore déterminée correspondant au feu, *iii)* le phlegme associé à un caractère réservé et lié à l'eau et *iv)* le sang, associé à l'air, considéré comme la source d'énergie du corps et de l'âme et à l'origine des natures enthousiastes et sociales. Pour Hippocrate, l'équilibre entre les humeurs émises par les différentes parties du corps est transmis héréditairement par les organes génitaux des parents au travers de la semence (« *semen* », la graine) et leurs proportions respectives déterminent la santé et le tempérament de l'enfant lors du mélange conduisant à la formation de l'embryon [6]. Cette théorie selon laquelle chaque partie du corps émettent continuellement son propre type de petite particule

se dresse comme les prémices du concept de « pangenèse » [7] que Charles Darwin mettra en lien avec la « théorie cellulaire » plusieurs siècles après. Quelques années plus tard, Aristote (384-322 av. J.-C.), qui a beaucoup écrit sur la médecine et la biologie tout en s'appuyant sur les travaux et les idées développées précédemment par Hippocrate, réfute la théorie des humeurs et propose une première théorie, centrée sur l'homme, qu'il viendra moduler par la suite, indiquant qu'il devait prendre en compte le rôle de la femme. Il propose ainsi, dans son ouvrage « *De la génération des animaux* », que l'homme apporte la « forme » (la semence) et que la femme apporte la « matière » nécessaire à la croissance de l'embryon [6, 8]. La dualité et la domination plus ou moins prononcées de cette forme sur la matière viennent alors, selon la nature de celles-ci, expliquées la transmission des caractères. Elles expliquent en partie pourquoi un enfant ressemble plus à son père ou à sa mère. Cependant, pour lui, « les fils ressemblent surtout au père et les filles à la mère » [9]. Dans la continuité de cette théorie, Aristote introduit toutefois le concept de « degré de parenté » en indiquant que l'enfant tend à ressembler davantage à ses aïeux proches qu'à ceux éloignés [9]. Notons que Platon (428-347 av. J.-C.) dans son ouvrage « *Timée* » avait déjà évoqué auparavant l'idée d'un mélange des semences parentales, masculine et féminine, influençant le devenir de la descendance, mais de façon beaucoup plus vague [10].

Ainsi, bien qu'ayant remarqué des ressemblances entre parents et enfants, Hippocrate et Aristote ne disposaient donc pas encore d'un modèle précis expliquant la transmission des traits héréditaires. Néanmoins, leurs écrits ont jeté les premiers fondements d'une réflexion rationnelle sur l'hérédité et l'origine des ressemblances familiales, dépassant les conceptions purement mythologiques ou religieuses précédemment établies.

1.1.2. Le gène, la « *brique* » de l'hérédité (de 1850 à 1940)

1.1.2.1. Les lois de Gregor Mendel au XIX^{ème} siècle : la naissance d'un modèle statistique discret de l'hérédité

Bien que les philosophes grecs aient jeté les bases d'une réflexion sur l'hérédité, il subsiste un vide de presque 2000 ans entre l'Antiquité et la découverte des lois de Mendel au XIX^{e} siècle. En effet, durant cette longue période, la majorité des naturalistes et philosophes sont soit restés tributaires de la théorie de la génération spontanée, soit cantonnés à des hypothèses du mélange inégal des semences. Ainsi, même si au XVII^{ème} siècle, Reinier de Graaf (1641-1673) et Antoni Van Leeuwenhoek (1632-1723), inventeur du microscope tel qu'on le connait aujourd'hui, ont permis d'observer d'une part les « œufs » des mammifères, chez les lapins en premier lieu [11], et d'autre part les spermatozoïdes [12], aucune supposition n'était faite quant à la complémentarité et à la contribution équivalente pour la descendance de ces deux éléments [6]. La théorie d'Aristote, considérant que l'un composait « la forme » et l'autre « le fond » restait alors en surface et permettait de répondre en partie à la complexité observée concernant la transmission des caractères d'une génération à une autre.

Ce n'est finalement qu'avec les travaux pionniers de Gregor Mendel (1822-1884) au milieu du XIX^{e} siècle, considéré aujourd'hui comme le « père de la génétique moderne », que le concept de « facteurs » (plus tard nommés allèles) ségrégeant lors de la reproduction et transmis intacts entre générations a pu être formulé. Malgré des premiers travaux sur les souris [13] ou encore les abeilles [14], ce sont ses expériences pionnières sur les pois, menées entre 1856 et 1863 – mais publiées en 1866 [15] – qui ont jeté les bases de la génétique moderne en instaurant les lois fondamentales régissant la transmission des caractères héréditaires dits caractères mendéliens.

Ainsi, Mendel, moine augustin autrichien également professeur de sciences naturelles passionné de botanique, entreprit de croiser méthodiquement différentes variétés de pois présentant des caractères (plus tard nommés phénotypes) contrastés pour sept caractères bien définis [16, 17] (forme et couleur des graines, forme et couleur des gousses, couleur et position des fleurs, taille des tiges). Après avoir établi des lignées qu'il considérait comme pures, il les croise entre-elles et s'intéresse à la transmission des caractères précédemment définis. Une des spécificités de ces travaux et que, s'il s'intéresse à ses plants d'un point de

vue qualitatif notamment en consignant méthodiquement leur apparence, il a aussi apporté une notion quantitative en quantifiant le nombre de plants exposant chacune des caractéristiques et ce pour un grand effectif. Les expériences sont en effet menées sur un ensemble de plus de 30 000 plants de pois [15, 18].

Ses croisements méticuleux lui permirent de tirer plusieurs conclusions desquelles sont formulées les trois lois fondamentales de l'hérédité [19, 20]. Ainsi, suite au croisement d'une lignée pure à fleurs violettes et d'une autre à fleurs blanches, il obtient une génération hybride fille (nommée F1) complétement violette. Ce premier résultat va déjà à l'encontre des croyances de l'époque qui prédisaient l'obtention de fleurs violettes-pâles ou blanches-violacées, résultat de « l'hérédité par mélange » [21]. C'est, par ailleurs, suite à cette première observation sur la lignée F1 que Mendel définit les caractères dits « dominants » (ici, la couleur violette) et les caractères « récessifs » (ici, la couleur blanche) qui est formulé dans l'une des lois :

1) ***La loi de la dominance*** – Lorsqu'un individu hérite de deux allèles différents d'un gène, seul l'un des deux s'exprime dans le phénotype.

Fort de ce constat, Mendel eut alors l'idée de croiser les descendants F1 entre eux, par autofécondation, afin d'obtenir une seconde génération, dites F2. Ce croisement résulte ensuite en l'obtention d'un quart de fleurs blanches et de trois quarts de fleurs violettes, obtenant un « ratio 3:1 ». Mendel établit ainsi une seconde loi permettant d'apporter des explications à la fois biologiques et statistiques à ce ratio :

2) ***La loi de la ségrégation*** – Chaque gamète (mâle et femelle) ne dispose que d'une seule copie d'un allèle initialement présent dans l'organisme. Suite à la fécondation, les gamètes fusionnent aléatoirement, l'embryon est alors composé de la somme des allèles précédemment contenue indépendamment dans chaque gamète.

Le tableau de croisement, résultat de la distribution des allèles, sera appelée échiquier de Punnett du nom de son inventeur éponyme Réginal Punnett (1875-1967) qui quelques années plus tard, lors de la redécouverte des lois de Mendel au début du XX^{e} siècle standardisera son style et son usage [22].

Si ces observations ont été, en premier lieu, faites à partir de l'observation de la couleur des fleurs, les mêmes conclusions ont pu être tirées à partir des autres caractères à l'étude par Mendel permettant de renforcer les lois évoquées.

Afin d'enrichir ces travaux, Mendel s'intéresse alors à la transmission simultanée des caractères. Pour ce point, il réalise des expériences de croisements « di-hybrides », *i.e.*, en considérant des parents différents pour deux caractères et donnant ainsi des générations F1 hétérozygotes pour deux caractères. Tandis qu'à l'issue des croisements mono-hybrides, un ratio 3:1 était obtenu, les croisements « di-hybrides » résultent en un ratio 9:3:3:1. Ce ratio correspond à ce qui serait théoriquement obtenu dans l'hypothèse d'une répartition aléatoire de chaque allèle de chaque caractéristique. La concordance entre les ratios observés et théoriques montre alors que chacun des deux allèles est hérité indépendamment de l'autre, avec un rapport phénotypique de 3:1 pour chacun. À la suite de ce résultat, Mendel énonce donc la dernière loi de l'hérédité :

3) ***La loi de l'assortiment indépendant*** – Les allèles de deux (ou plusieurs) gènes différents sont répartis en gamètes indépendamment les uns des autres

Si les travaux de Mendel sonnent de nos jours comme une révolution scientifique, ces travaux ont à l'époque été ignorés par les scientifiques, tout du mois d'un point de vue de l'hérédité et pour leur aspect génétique, et ont même pu être sévèrement remis en cause notamment par les défenseurs de « l'hérédité par mélange ». Ronald Fisher (1890-1962) par exemple recalculera les ratios trouvés par Mendel en remettant en cause les conclusions tirés [23]. Charles Darwin (1809-1882), suite aux travaux de Jean-Baptiste de Lamarck (1744-1828), quant à lui proposera sa propre théorie « pangénique » de l'hérédité dans laquelle il propose que chaque partie du corps émette continuellement son propre type de petites particules organiques apportant des informations héréditaires aux gamètes [7].

Cette compétition entre différentes théories aura pour conséquence de mettre sur le côté les travaux de Mendel, qui devront attendre le début du XX^{e} siècle pour être remis en lumière.

Ainsi, si les travaux de Mendel n'ont pas tout de suite été reconnus, ils jetèrent les bases d'un modèle discret de l'hérédité reposant sur des facteurs transmis intacts entre générations selon des rapports prévisibles statistiquement. Mendel posa ainsi les jalons du concept de gène, bien que le terme ne fût pas employé *stricto sensu* dans ses travaux. Ses lois ouvrirent

la voie à la compréhension des mécanismes sous-jacents régissant l'hérédité, phénomène jusqu'alors observé, mais non expliqué par des théories mathématiques.

1.1.2.2. La redécouverte des lois de Mendel au début du XXème siècle : Loi de l'hybridation et considération du chromosome comme support de l'information génétique

Si les travaux de Mendel sont restés dans l'ombre pendant plusieurs décennies, ce sont les travaux de trois botanistes européens Carl Correns (1864-1933), Hugo De Vries (1848-1935) et Erich Von Tschemark (1871-1962), qui vont permettre de rebattre les cartes. En effet, De Vries et Correns auraient, sans connaissance de l'article de Mendel, développé des conclusions et des hypothèses très proches qui soulignaient les mêmes principes que ceux que Mendel avait théorisés [24, 25]. Concernant Tschermak, quoiqu'il semble avoir compris l'importance du travail de Mendel, ces travaux étaient plus litigieux quant aux conclusions tirées [26].

Notons que De Vries est le premier à employer le terme « pangène » (inspiré de la pangenèse de Darwin, mais rejetant la vision Lamarckienne) pour définir des particules transmises, lors de la division cellulaire, aux cellules filles. Ils supposent que ces particules sont conduites aux cellules germinales et transmettent les caractères que les cellules respectives ont pu acquérir au cours du développement [24, 27]. De plus, il est à l'origine de la théorie de la mutation, considérant que ces dernières sont nécessaires à l'apparition de nouveaux caractères (et ainsi de nouvelles espèces).

En parallèle de ces travaux portant sur l'hérédité et sur les règles sous-jacentes, les prémices de la biologie moléculaire se développent. En effet, en 1871, Friedrich Miescher (1844-1895) est le premier à isoler de l'ADN à partir de lymphocytes [28]. Durant les années 1880, Walther Flemming (1843-1905) observe et définit la chromatine et constate que cette structure est associée à des structures filiformes dans le noyau de la cellule [29, 30], les chromosomes tels que nommés par Heinrich Wilhelm Waldeyer (1836-1921) [31, 32] quelques années plus tard. De manière indépendante, Edouard Van Beneden (1846-1910) avait également observé de telles structures. Finalement, ils observent et analysent conjointement la distribution des chromosomes pendant la division cellulaire que Flemming nommera à ce moment-ci

« mitose » [33]. Ainsi, dès 1900, les chromosomes étaient déjà connus et leur rôle en tant que support des facteurs héréditaires théorisés par Mendel est très vite considéré.

Ce lien postulé entre les gènes et les chromosomes, qui sera plus tard connu sous le nom de théorie chromosomique de l'hérédité, a été initialement établi par le biologiste allemand Theodor Boveri (1862-1915) et le généticien américain Walter Sutton (1877-1916) au cours des années 1902-1903 [34, 35]. En effet, par des observations microscopiques sur l'oursin, Boveri a d'abord démontré l'importance des chromosomes dans le développement embryonnaire [36]. Dans la continuité de ces travaux et se basant sur les premières observations de de Carl Rabl (1853-1917) concernant la continuité des chromosomes durant le cycle cellulaire, il développe la théorie de l'individualité des chromosomes indiquant que ces derniers sont gardiens de l'information génétique [37]. Sutton, pour sa part et sur la base de ses études sur la spermatogenèse des sauterelles, a démontré que les chromosomes fonctionnent par paire, composée d'un chromosome paternel et d'un chromosome maternel qui se séparent à la méiose [38]. L'hypothèse selon laquelle le comportement des chromosomes lors des divisions méiotiques pourrait alors être le support des lois de l'hérédité de Mendel se met ainsi en place. Si ces travaux permettaient de faire le lien entre les observations de Mendel et les chromosomes, ils restent cependant encore très controversés. Plusieurs travaux complémentaires, tels que ceux de Nettie Stevens (1861 - 1912), à l'origine de la découverte des « chromosomes sexuels » notamment en travaillant sur la détermination du sexe chez le ver de farine [39, 40] ou ceux d'Eleanor Carothers (1882 – 1957) chez le criquet [41] venant appuyer à nouveau les observations de Mendel sur l'assortiment indépendant des chromosomes seront nécessaires pour renforcer la théorie de l'hérédité par chromosome, mais ils ne seront pas suffisants pour lever les controverses. Durant cette période charnière et plus précisément aux alentours de 1909, Wilhelm Johannsen (1857-1927) instaure les termes « gène », « génotype » et « phénotype » [42]. De manière intéressante, au moment même de la création du terme « gène », le concept semblait ne pouvoir être défini de manière précise [43].

Si cette période permet de faire le lien entre les lois de Mendel et les chromosomes comme support de l'information héréditaire, il faudra attendre les travaux de Thomas Hunt Morgan (1866 -1945) entre 1910 et 1930 venant appuyer le concept de gêne pour que cette théorie se présente comme difficilement réfutable.

1.1.2.3. Les travaux pionniers de Morgan et son équipe sur la drosophile : de la découverte des gènes liés aux chromosomes à l'élaboration des premières cartes génétiques

Les travaux pionniers de Thomas Hunt Morgan (1866-1945) sur la drosophile au début du XX^{e} siècle jetèrent les bases de la génétique moderne en établissant le rôle des chromosomes dans l'hérédité et la nature discrète des gènes. En effet, en utilisant la mouche du vinaigre comme organisme modèle de par son cycle de vie court, ses nombreux descendants et ses divers mutants naturels aux phénotypes visibles, Morgan, avec l'aide de ses étudiants, purent cartographier les premiers gènes sur les chromosomes et démontrer que les gènes sont transmis en bloc le long des chromosomes. Ainsi, si les premières années d'expérimentation et de croisement n'étaient pas concluantes du fait à la fois d'un nombre élevé d'individus qui s'échappaient et se mélangeait facilement, mais également de mutants avec des effets phénotypiques subtils, à partir de 1909, Morgan et son équipe commencent à observer des mutants héritables venant appuyer de nouveau les lois de l'hérédité de Mendel. En 1911, observant la couleur des yeux chez ses drosophiles, Morgan remarque un mâle mutant aux yeux blancs alors que les autres membres de la population, mâles comme femelles, présentent les yeux rouges. Suivant les expériences de Mendel, il décide de faire accoupler les mouches mâles aux yeux blancs avec des femelles aux yeux rouges et obtient des hybrides (F1) uniquement aux yeux rouges. Dans la continuité, il réalise des croisements de deuxième ordre (F2) où il n'obtient que des mâles aux yeux blancs, ce qui allait à l'encontre des ratios mendéliens. Cette expérience permit à Morgan d'identifier le premier gène lié au sexe, le gène *white* qui détermine la couleur des yeux chez la drosophile. De cette expérience, il conclura que certains caractères sont liés au sexe de par l'existence des chromosomes sexuels et hypothétisera que d'autres gènes sont probablement également portés par des chromosomes spécifiques [44, 45]. Dans la lignée de ces analyses, Morgan, son équipe et notamment Alfred Sturtevant (1891-1970) commencèrent à réaliser de nombreux croisements en suivant plusieurs caractères afin d'étudier les lois de l'hérédité de manière plus exhaustive [46, 47]. Ainsi, ils identifièrent le gène lié à la taille des ailes (*miniature-wing*) et montrèrent qu'il était, lui aussi, lié au sexe. Cependant, réalisant de nombreux croisements, ils observèrent que l'hérédité de ce caractère était indépendante du caractère lié aux yeux blancs. C'est suite à ces analyses qu'ils proposeront les concepts de « déséquilibre de liaison » entre gènes et de « *crossing over* » ainsi que l'idée de distance séparant les gènes sur le chromosome. C'est par

ailleurs de cela que naitra la métrique « (centi)morgan » permettant de mesurer la liaison génétique entre gènes. Partant donc de cette idée que les gènes sont localisés de façon linéaire et précise le long des chromosomes, Alfred Sturtevant proposera en 1913 la première carte de liaison génétique chez la drosophile.

À partir de ces travaux et de l'expansion de ce type d'analyses et d'observations, la théorie chromosomique de Boveri-Sutton fut de plus en plus accepté par la communauté scientifique même si des débats restaient en suspens notamment sur l'interaction entre les gènes ou encore sur le concept même du gène, certains pensant qu'il n'y avait aucune raison valable de considérer le gène comme une unité discrète de l'hérédité.

1.1.3. L'ère de l'ADN : approche moléculaire du gène (1930 – 1970)

1.1.3.1. Des enzymes aux acides nucléiques : l'avènement de la biologie moléculaire et l'émergence du dogme central posant la définition classique du gène (1930-1965)

Les années 1930 marquent l'avènement de la biochimie des enzymes, grâce à l'utilisation de nouvelles techniques de fractionnement et de purification comme l'ultracentrifugation et l'électrophorèse. Les bases de la biologie moléculaire se mettent ainsi en place avec les travaux notamment de James Sumner (1887-1955), John Northrop (1891-1987) et Wendell Stanley (1904-1971). En 1926, Sumner parvient à isoler et à cristalliser pour la première fois une enzyme, l'uréase, démontrant ainsi sa nature protéique [48]. Quelques années plus tard, Northrop et Stanley isolent la pepsine et la trypsine, confirmant par la suite le fait que les enzymes soient bien des protéines [49, 50]. Ces découvertes jetèrent les bases d'une approche biochimique et moléculaire du fonctionnement cellulaire et de l'étude des gènes. Dans les années 1940, la piste de l'acide désoxyribonucléique (ADN) comme support de l'information génétique se précise. En 1928, Frederick Griffith (1879-1941) avait déjà démontré l'existence d'un "principe transformant" transmissible lors de ses expériences sur la bactérie *Streptococcus pneumoniae* [51]. Ce n'est qu'en 1944, en utilisant des méthodes de transformation bactérienne, que Oswald Avery (1877-1955), Colin MacLeod (1909-1972) et Maclyn McCarty (1911-2005) parviennent à identifier ce principe transformant comme étant l'ADN, fournissant la première preuve que les gènes sont portés par les molécules d'ADN et non par les protéines elles-mêmes [52]. Cependant, ces travaux peinent à convaincre la

communauté scientifique, encore sceptique sur le rôle de l'ADN [53]. Sa fonction comme support universel de l'information génétique est finalement établie dans les années 1950. En 1952, Alfred Hershey (1908-1997) et Martha Chase (1927-2003) utilisent des isotopes radioactifs pour marquer sélectivement l'ADN et les protéines du bactériophage T2 et démontrent ainsi que c'est l'ADN, et non les protéines, qui pénètre dans les cellules infectées, apportant la preuve définitive du rôle de l'ADN [54]. En 1953, Rosalind Franklin (1920-1958) et Maurice Wilkins (1916-2004) observent pour la première fois, à partir de données de diffraction des rayons X (le cliché 51), la structure de l'ADN. Ce servant de ces travaux, James Watson (né en 1928) et Francis Crick (1916-2004) proposent alors un modèle en double hélice pour qualifier la structure de l'ADN, scellant ainsi le statut de celui-ci comme matériel génétique [55, 56]. En effet, sur la base de cette structure tridimensionnelle, Watson et Crick proposent ainsi un modèle de réplication semi-conservative de l'ADN. Ces découvertes mènent à la formalisation du "dogme central" de la biologie moléculaire par Francis Crick en 1958, stipulant que l'information génétique contenue dans l'ADN est transférée vers l'ARN messager, puis traduite en protéines [57]. Ce dogme est alors la base du cadre conceptuel de la définition classique du gène, considéré suite à ces travaux comme une séquence d'ADN codant pour une protéine. Durant cette période, les travaux sur le code génétique menés par Marshall Nirenberg (1927-2010), Johann Matthaei (né en 1929), Har Khorana (1922-2011) et leurs collègues permettent également de comprendre le lien entre l'information portée par l'ADN et la synthèse des protéines. Dès 1961, Nirenberg et Matthaei réalisent une expérience clé en incubant un ARN messager poly-uracile avec un système acellulaire de synthèse des protéines. Ils obtiennent exclusivement la synthèse de poly-phénylalanine, démontrant que le codon UUU code pour la phénylalanine [58, 59]. Par la suite, Khorana synthétise des homopolymères d'ARN avec des séquences répétées de chaque nucléotide. Combinés au système acellulaire, ces homopolymères permettent d'identifier les acides aminés correspondant aux codons composés uniquement de A, U, G ou C [60]. En 1965, le code est *cracké* quasi complètement en combinant ces approches avec de nouvelles techniques biochimiques comme la synthèse d'ARN messagers de séquence définie et la séparation des acides aminés marqués sur chromatographie. Il est ainsi démontré que le code génétique est dégénéré, redondant, mais non ambigu.

Ainsi, entre les années 1930 et 1950, les bases de la biologie moléculaire sont établies, avec l'identification biochimique des enzymes, la démonstration du rôle de l'ADN comme support

universel de l'information génétique, et l'élucidation de la structure de l'ADN et du dogme central. Ces découvertes fondatrices ont permis de jeter les bases d'une définition classique du gène, cantonné à sa séquence codante d'ADN résultant en protéine.

1.1.3.2. Le concept de gène remis en question : l'impact des nouvelles techniques de biologie moléculaire sur la conception classique du gène (années 1960-1970)

À partir des années 1960, le développement de nouvelles techniques de biologie moléculaire commence à ébranler la définition classique du gène comme simple séquence d'ADN codante. L'avènement des méthodes de séquençage comme la dégradation partielle et la méthode de Sanger dans les années 1970 permet d'analyser plus finement l'organisation des gènes. Le clonage moléculaire d'ADN complémentaire couplé aux prémices des analyses bio-informatiques a également montré l'existence de séquences interrompant les gènes, les introns, qui sont épissées pour donner la séquence codante finale, les exons [61, 62]. Ce phénomène d'épissage alternatif complexifie la relation univoque entre la séquence d'ADN et la séquence protéique. De plus, l'étude des génomes procaryotes par des approches de traduction *in vitro* met en évidence des gènes qui se chevauchent, voire imbriquées. Un même segment d'ADN peut ainsi coder pour plusieurs protéines distinctes, remettant en cause le concept d' « un gène une protéine ». La découverte des rétrovirus et de leur enzyme clé, la transcriptase inverse, bouleverse par ailleurs le dogme central [63, 64]. Ces virus utilisent un intermédiaire ARN pour rétro-transcrire leur matériel génétique sous forme d'ADN et l'intégrer dans le génome de leur hôte. Ce flux d'information inverse remet en question le caractère unidirectionnel du dogme central et souligne sa nature plus complexe qu'initialement envisagée. Par ailleurs, les analyses biochimiques révèlent le rôle majeur des éléments non-codants dans la régulation de l'expression des gènes. En 1961, François Jacob (1920 – 2013) et Jacques Monod (1910 – 1976) décrivent le modèle de l'opéron lactose, montrant que des protéines régulatrices se fixent sur des séquences opératrices en amont des gènes pour contrôler leur transcription [65, 66]. Ces éléments régulateurs situés localement ou à distance et agissant directement (-*cis*) ou par l'usage d'un intermédiaire (-*trans*) ajoutent un niveau de complexité supplémentaire absent de la théorie classique du gène. Enfin, les progrès dans l'étude des ARN non-codants conduisent à reconnaître leur importance

fonctionnelle. Les ARN de transfert et ARN ribosomiques participent activement à la synthèse protéique [67, 68], tandis que les petits ARN nucléaires guident l'épissage des pré-ARNm. Plutôt que de simples intermédiaires, les ARN sont désormais vus comme des acteurs clés de l'expression génique. Ainsi, à partir des années 1960, le développement de nouvelles approches expérimentales en biologie moléculaire a profondément remis en cause la conception classique du gène, en révélant sa complexité tant au niveau structurel que fonctionnel.

1.1.4. Le gène à l'ère post-génomique (depuis les années 1990)

1.1.4.1. Démocratisation du séquençage haut débit à grande échelle et caractérisation des éléments « non-codants »

À partir des années 1990, le séquençage et l'analyse à grande échelle des génomes complets, permises par des approches de séquençage haut-débit comme le séquençage à haut-parallélisme d'Illumina, conduisent à une réévaluation fondamentale de la définition du gène et de son rôle au sein du génome. En 2001, la finalisation du séquençage du génome humain marque un jalon historique dans l'accès à l'intégralité de l'information génétique d'un organisme [69, 70]. Parallèlement, autour des années 2000, de nombreux autres génomes modèles sont séquencés, comme la levure [71], la drosophile [72], le nématode *C. elegans* [73] ou la plante *A. thaliana* [74]. Le premier génome concernant les animaux d'élevage sera par ailleurs celui de la poule séquencée en 2004 [75]. En effet, l'effort réalisé pour cette espèce est en lien avec l'importance de celle-ci dans différents secteurs. En 2021, la volaille (majoritairement la poule) est la viande la plus consommée dans le monde avec environ 138 millions de tonnes produites à l'année et plus de 92 millions de tonnes d'œuf [76]. Cela représente un cheptel d'approximativement 27,5 milliards de poules à l'année. De plus, les prévisions et les données sur ces dernières années tendent à montrer une consommation croissante, la poule étant globalement non soumise à des restrictions culturelles ou religieuses. Elle présente également un coût de production plus faible et est globalement moins consommatrices de ressources [77]. Sur le pan de la recherche, la poule se présente comme un modèle d'étude aux multiples avantages. En effet, c'est un animal de petite taille, demandant un entretien faible par rapport aux autres espèces et avec un temps de génération

court (généralement moins d'un an) favorisant les études. La poule est notamment utilisée dans le cadre de travaux sur le développement des vertébrés, car l'accès à l'embryon se fait facilement [78]. Pour finir, sa divergence phylogénétique avec l'humain datant d'environ 300 millions d'années permet les études de conservation durant l'évolution [79].

Ainsi, l'ensemble des intérêts de différents ordres ont permis de générer d'importantes quantités de données génomiques. Ces données globales mettent en évidence que les régions codantes, *i.e.* les codons à l'origine des acides aminés, représentent souvent moins de 2% du génome chez les eucaryotes [80, 81]. La majeure partie du génome est ainsi constituée de séquences non-codantes, nuançant l'importance accordée précédemment au gène défini par sa séquence codante. De plus, on découvre que l'environnement et les conditions externes ont un impact majeur sur l'expression des gènes, via notamment des mécanismes épigénétiques comme la méthylation de l'ADN et les modifications d'histones, modulant l'état de la chromatine [82]. Selon les conditions environnementales, le même gène peut être activé ou réprimé via ces mécanismes épigénétiques. À ce moment, le gène n'est plus une entité fixe, imperméable, mais une séquence dont l'expression est modulable en fonction du contexte cellulaire et physiologique. Les approches de génomique fonctionnelle se développent afin d'étudier les interactions entre les gènes. Les puces à ADN et le séquençage ARN permettent d'analyser l'expression de milliers de gènes en parallèle [83–85]. Il apparait alors que les gènes font partie de réseaux d'expression complexes, leur expression étant coordonnée. La définition du gène s'inscrit désormais dans une vision plus globale et systémique du génome. Enfin, les éléments mobiles, longtemps considérés comme de l'ADN "égoïste", sont reconnus comme des acteurs clés de l'évolution et de la plasticité du génome des eucaryotes [86–88]. Ces séquences, capables de se déplacer et de se multiplier de façon autonome, représentent plus de 40 % du génome humain. Elles participent aux réarrangements génomiques et à la régulation de l'expression des gènes adjacents.

Ainsi, l'avènement de la génomique bouleverse profondément la conception du gène, en révélant l'importance majeure du contexte génomique, épigénétique et réseau dans lequel il s'inscrit. La définition du gène tend à s'élargir pour englober toutes ces dimensions d'expression.

1.1.4.2. Bio-informatique et génomique : un tandem indissociable pour repenser le gène et le génome

L'avènement de la bio-informatique à partir des années 1970 a fourni des outils essentiels pour donner du sens aux données massives générées par la génomique, transformant notre compréhension du génome et du gène. Le terme de bio-informatique est mentionné pour la première fois en 1970 par Ben Hesper (*information indisponible*) et Paulien Hogeweg (née en 1943) pour désigner l'application de l'informatique à la biologie [89, 90]. La bio-informatique peut alors être définie comme la science consistant à stocker, organiser, analyser et intégrer les informations biologiques à l'aide d'approches computationnelles et statistiques. Les prémices de cette discipline remontent aux années 1970 avec le développement des premiers algorithmes pour analyser des séquences biologiques, citons par exemple l'alignement de séquences par Needleman-Wunsch [91]. C'est notamment avec l'émergence des techniques de séquençage et de clonage de l'ADN, que naissent les premières banques de données de séquences nucléiques et protéiques, comme GenBank produit au sein du NIH (*National Institute of Health*, États-Unis) branche du NCBI (*National Center for Biotechnology Information*, États-Unis) ou encore la base de données du EMBL (*European Molecular Biology Laboratory*) maintenu par l'EBI (*European Bioinformatics Institute*) [92, 93]. Aux débuts des années 1980, les méthodes d'alignement de séquences permettant de rechercher des similitudes avec les banques de données se développent et différents algorithmes avec chacun leur spécificité voient le jour comme BLAST [94] ou FASTA [95]. Ces bases de données et outils bio-informatiques se révèlent indispensables pour stocker et analyser les premiers génomes séquencés et prennent toute leur importance avec les projets de séquençage de génomes complets variés. L'assemblage et l'annotation de ces génomes nécessitent des approches bio-informatiques à grande échelle pour construire des cartes génomiques et identifier les gènes. Des algorithmes de prédiction de gènes sont ainsi mis au point pour détecter les régions codantes sur la séquence nue, comme Genscan [96, 97], et se révèlent cruciaux pour l'annotation des premiers génomes séquencés. La comparaison des séquences permet également d'analyser l'évolution des génomes par des méthodes comme le calcul de distance génétique. Les outils bio-informatiques permettent par ailleurs de mettre en évidence les éléments répétés, mobiles, ainsi que l'importance des séquences non-codantes dans les génomes séquencés, comme évoqué précédemment. Ils sont essentiels pour intégrer les données de transcriptomique et épigénomique, révélant la complexité de la régulation de

l'expression des gènes. L'analyse des séquences à haut-débit issue des nouvelles technologies de séquençage est ainsi cruciale pour l'identification des micros ARN, des ARN longs non codants et autres éléments non-codants. Elle constitue le pilier des approches de génomique comparative visant à élucider l'évolution des génomes.

Ainsi, sans le support de la bio-informatique, l'énorme masse de données génomiques générées ces dernières décennies serait inexploitable. Ces outils ont transformé notre appréhension du génome et la définition du gène en permettant d'en quantifier la complexité.

1.2. Les ARN longs non-codants : un nouveau type de gène impliqué dans la régulation de l'expression génique et des phénotypes complexes *(Revues collaboratives 1 & 2)*

Ce chapitre est en partie inspiré de deux revues qui ont été réalisées durant la thèse.

La première (référencée dans le paragraphe qui suit comme « *review-1 »)* fournit une vue générale concernant les ARN longs non codants (lncRNA) notamment dans les espèces d'élevages telles que le bovin, le porc, la volaille, l'équin et le chien tout en les analysant conjointement aux espèces de référence que sont l'humain et la souris. L'analyse simultanée de l'ensemble de ces espèces permet de couvrir une large échelle phylogénétique. Si les méthodes employées usuellement pour l'étude des lncRNA sont présentées, les notions en lien avec la génomique comparative sont également discutées dans l'objectif de renforcer l'annotation de ces lncRNA. Pour finir, un ensemble de lncRNA mis en relation avec des traits/phénotypes spécifiques par le biais d'études d'association sont décrits ainsi que des stratégies alternatives visant à augmenter le petit nombre de lncRNA validés sur le plan fonctionnel chez les animaux domestiques.

Ce travail a donc fait l'objet :

- d'une publication : Lagarrigue S, Lorthois M, **Degalez F**, Gilot D, Derrien T (2022). LncRNAs in domesticated animals: from dog to livestock species. Mammalian Genome. doi: 10.1007/s00335-021-09928-7. **Cet article est reproduit en fin de §1.2.**

La seconde revue (référencée dans le paragraphe qui suit comme « *review-2 »)* s'intéresse à l'évolution des assemblages du génome de la poule, depuis le premier, galgal2, paru en 2004 jusqu'aux derniers faisant référence GRCg7b et GRCG7w de juin 2022. Les travaux portent également sur l'évolution des annotations du génome associées à ces assemblages et qui sont fournies par les deux bases de données de référence Ensembl et RefSeq. En effet, en 2015, avec l'introduction des gènes lncRNA dans les annotations, le nombre de modèles identifiés a connu une croissance significative, même si la liste est loin d'être exhaustive. Ainsi, les divergences d'annotations entre ces deux sources ont été étudiées et soulignent la méconnaissance des modèles de type lncRNA. Pour finir, deux initiatives visant à combiner plusieurs sources d'annotations sont présentées et discutées.

Ce travail a donc fait l'objet :

- d'une publication : **Degalez F,** Muret K, Lagarrigue S (2023). Evolution of protein coding and long non coding genes of the chicken genome through the different genome assemblies and their associated annotations. Cytogenetic and Genome Research. doi: 10.1159/000529376. Ce travail s'inscrit dans le cadre d'un travail collaboratif pour le « Fourth Report on Chicken Genes and Chromosomes 2022 » mené par la docteure Jacqueline Smith de l'Université d'Édimbourg en Ecosse. ***Seule la partie correspondant à la présentation globale du travail par J. Smith ainsi que celle ayant trait à ce qui est présentée en amont sont reproduites en fin*** **de §1.2.**

Même si ces travaux constituent des travaux propres à la thèse, ils ne sont pas réintroduits dans la partie « Articles et travaux complémentaires ». Outre les synthèses exposées ci-dessus, aucun résumé de l'article n'est alors fourni, mais une grande partie des propos, figures et exemples utilisés par la suite se réfèrent, comme indiqué, à ces documents.

1.2.1. Prédiction des lncRNA et estimation de leurs profils d'expression par approches bio-informatiques

Les lncRNA sont une classe d'ARN transcrits par l'ARN polymérase II, de plus de 200 nucléotides de longueur, dépourvus de cadre ouvert de lecture significatif et qui ne codent donc pas pour des protéines [98–100]. Leur définition s'est précisée au cours des années, notamment grâce aux avancées des technologies de séquençage à haut-débit et continue d'être sujet à débat [101, 102]. En effet, l'avènement du séquençage ARN dans les années 2000 a permis d'identifier de nombreux transcrits jusque-là inconnus dans le génome, transcrits qui se sont révélés majoritairement non codants [99, 103–107]. En effet, si dans les bases de données de référence et quelle que soit l'espèce [103], le nombre de gènes codant des protéines (PCG pour *Protein Coding Gene*) avoisinent les 20 000, le nombre de lncRNA prévisionnels reste sujets à débat, les expressions s'approchant du « bruit transcriptionnel » [108–111]. Ainsi, le nombre de lncRNA identifiés dans ces bases de données de référence est très fluctuant selon les sources et les espèces, vacillant entre 1 000 et 17 000 [103] selon l'espèce et les efforts d'annotation associés, mais les prédictions font état de plusieurs dizaines/centaines de milliers de lncRNA [112, 113]. Dans tous les cas, ces lncRNA forment une classe très hétérogène, et leurs fonctions restent encore méconnues dans de nombreux cas. Cependant, pour ceux finement étudiés, ils apparaissent agir sous forme d'ARN via des interactions avec des protéines, des ARN ou encore de l'ADN génomique [100, 103, 114, 115]. Une description plus détaillée des fonctions de ces lncRNA est proposée dans le §1.2.3.1.

La prédiction des lncRNAs repose principalement sur des approches bio-informatiques, utilisant à la fois les données de séquençage et les informations de génomique computationnelle. Tout comme pour les gènes codant des protéines, les lectures (appelées *reads*) issues du séquençage ARN sont tout d'abord alignées sur le génome de référence en considérant les jonctions exon-intron afin de reconstruire les transcrits (voir §2.1.1). Ensuite, des outils de reconstruction de transcrits dédiés tels que Cufflinks [116] ou StringTie [117] utilisent les *reads* précédemment alignés dans le but de cartographier sur le génome, soit des transcrits connus, c'est-à-dire déjà présents dans l'annotation de référence, soit de nouveaux transcrits (voir §2.1.2). Spécifiquement pour les lncRNA, une série de filtres variant selon les approches, mais évaluant leur potentiel codant, est appliquée pour distinguer les lncRNA putatifs des autres transcrits codants ou non-codants. Plus précisément, des outils comme

CPC2 [118] analysent la structure des ORF (*Open Reading Frame* potentiels – Cadre de lecture ouvert), des parties d'une séquence d'ADN ou ARN regroupée en triplets de nucléotides consécutifs et susceptibles ou non d'être traduit en protéine ou en peptide. D'autres, comme PhyloCSF [119], examinent des signatures évolutives caractéristiques des alignements de régions codantes conservées, telles que les fréquences élevées des substitutions synonymes de codons et des substitutions conservatrices d'acides aminés (*synonymous variant*), et les faibles fréquences des autres substitutions faux-sens (*missense variant*), et non-sens (*stop gained*). De plus, des comparaisons peuvent être faites avec des bases de données de protéines connues comme UniProt/Swiss-Prot [120] pour éliminer les transcrits présentant des similarités. D'autre part, des programmes tels que PLEK [121] ne considère pas les ORF mais se base sur la fréquence de k-mers, des petites portions de séquences allant généralement de 1 à 10 nucléotides. D'autres outils comme FEELnc [122] combinent plusieurs de ces approches et incorporent également des données génomiques sur la structure des gènes pour affiner la prédiction [103].

Par la suite, en utilisant les modèles de gènes modélisés ou ceux décrits dans les bases de données de référence, l'expression peut par ailleurs être analysée (voir §2.1.2). Contrairement aux PCG qui sont fréquemment exprimés à des niveaux élevés, les lncRNA présentent souvent des niveaux d'expression beaucoup plus faibles et plus variables [103, 108, 123–125]. En effet, comme montré dans la *review-1* (Figure 2A) et en utilisant quatre tissus différents (tissue adipeux, foie, sang, hypothalamus), il apparait que plus de 90 % des *reads* s'alignent sur des PCG et que 80 % des *reads* correspondent aux 25 % des PCG les plus exprimés [103]. Par conséquent, peu de *reads* sont associés aux lncRNA ce qui explique en partie la difficulté à les identifier et à les caractériser finement et de manière redondante au sein de diverses analyses ou bases de données, justifiant ainsi la mise en place de ressources agrégeant les modèles géniques. En plus de leurs niveaux d'expression modestes, les lncRNA sont très dépendants des échantillons observés et des conditions dans lesquels ils ont été analysés (conditions-spécifiques), deux facteurs pouvant fluctuer, notamment pour les bases de données de référence. Ainsi, la comparaison des annotations du génome de la poule provenant de Ensembl et de Refseq montre un faible chevauchement (voir *review-1* Figure 2B et *review-2* Figure 13), à hauteur d'environ 14 % pour les modèles de gènes de type lncRNA, contre plus de 87 % pour les PCG [103]. En effet, de nombreux lncRNA (plus que pour les PCG) présentent par exemple une expression tissu-spécifique, c'est-à-dire qu'ils sont exprimés

préférentiellement dans certains tissus et/ou types cellulaires [124, 125]. Par exemple, le lncRNA TINCR est connu chez l'humain pour être spécifique de la peau et des tissus kératinisés [126]. De même, chez la poule, le gène KRT75L4 référencé dans la base de donnée OMIA [127] comme responsable du frisage des plumes [128, 129] apparait uniquement exprimés dans ce tissu [125]. Cette spécificité tissulaire est importante, car elle suggère des fonctions biologiques en lien avec le tissu considéré, ce qui doit être considéré lors de la mise en place des études et des plans expérimentaux.

De la même manière, certains lncRNA présentent également une expression temporelle spécifique. Leur expression peut varier au cours du développement, de la différenciation cellulaire ou en réponse à des stimuli externes. Un des exemples les plus connus est celui du lncRNA Coolair qui est activé spécifiquement au cours de la floraison chez *Arabidopsis thaliana* [130], cependant d'autres mécanismes de ce type ont été identifiés tel que le lncRNA HOTAIR qui présente une expression régulée au cours de la différenciation des fibroblastes chez l'humain [131]. L'observation de deux lignées de poule divergentes pour la lipogenèse et challengées pour deux types d'alimentation, dont l'une était supplémentée en fibres, a permis d'identifier de nombreux gènes différentiellement exprimés pour l'alimentation [132]. Par exemple, le lncRNA « LOC107053670 », convergeant avec le gène codant la protéine SCD (steroyl-CoA desaturase) affichait une expression spécifique dans le foie et en réponse au régime alimentaire (données personelles). Ces variations d'expression temporelles indiquent par ailleurs que ces lncRNA sont impliqués dans la régulation de processus dynamiques.

L'intégration de toutes ces données permet globalement de proposer des catalogues de lncRNA prédits pour un organisme, un tissu, et/ou un type cellulaire donné. Cependant, seule une petite fraction de ces lncRNA prédits *in silico* sont validés et caractérisés fonctionnellement. Les faux positifs doivent notamment être éliminés expérimentalement. De plus, ces approches informatiques présentent certains biais, comme une sous-estimation des lncRNA à faible niveau d'expression ou une surestimation de ceux chevauchant, sur le même brin, des régions codantes [99] lorsque ces derniers sont conservés dans les annotations.

1.2.2. Classifications des longs ARN non-codants : aide pour leur caractérisation et l'investigation de leurs mécanismes d'action

Tout comme pour les gènes codants des protéines, établir des classifications précises et harmonisées des lncRNA est essentiel pour bien comprendre leur diversité et appréhender leurs fonctions. En effet, selon les classifications et les classes auxquelles ces gènes sont associés, il est possible d'émettre des hypothèses plus en moins fortes en termes de mécanismes d'action, de processus biologiques régulés et de localisation cellulaire.

Par exemple, la position et la configuration d'un lncRNA par rapport au PCG le plus proche peut permettre de supputer des mécanismes de régulation transcriptionnelle ou post-transcriptionnelle. Dans le cadre de cette classification, facilement accessible par l'application du module « classifier » de FEELnc, principalement deux catégories, elles-mêmes subdivisées en plusieurs sous-catégories peuvent être identifiées [122, 133, 134] :

- Les lncRNAs intergéniques (notés iRNA ou plus usuellement lincRNA), qui sont localisés entre deux gènes codants et ne chevauchent aucun gène connu. Parmi les plus étudiés, on connait historiquement, chez l'homme et la souris, le lincRNA XIST qui contrôle l'inactivation du chromosome X [135, 136]. Au sein même de cette catégorie, il est usuellement possible de distinguer trois sous classes selon leur configuration avec le PCG le plus proche :
    - Les convergents où les régions 5'UTR de chaque gène se font faces ;
    - Les divergents où les régions 3'UTR de chaque gène se font faces ;
    - Les même-brin (*same_strand*) où le lncRNA est dans le même orientation que le PCG.

- Les lncRNAs géniques (notés parfois gRNA), qui présentent une partie chevauchante avec un élément du PCG. Afin d'annoter plus finement ces catégories, trois critères complémentaires sont considérés :
    - L'orientation relative des gènes entre eux, ainsi si le lncRNA et le PCG sont portés par le même brin, ils sont considérés comme sens, dans le cas inverse, ils sont considérés comme antisens. Le cas des gènes sens est particulier, en effet, il ne s'agit pas d'un nouveau locus génique au sens strict, les exons se chevauchant, mais bien d'un transcrit pouvant être qualifié d'alternatif.
    - Le type d'élément qui est chevauché par les parties exoniques du lncRNA, ainsi, on distingue les lncRNA avec chevauchement intronique ou exonique.
    - La nature du chevauchement, ainsi le lncRNA peut être en partie chevauchant (*overlappant*), être niché intégralement dans un PCG (*nested*) ou alors contenir intégralement le PCG (*contained*).

À titre d'exemple, citons ANRIL (CDKN2B-AS1), antisens de CDKN2B, qui, chez l'homme, régule l'expression de ce dernier [137], ou UCHL1-DT (aussi nommé UCHL1-AS1) qui contrôle la traduction de la protéine UCHL1 [138]. Ces deux cas peuvent ici être considérés comme des lncRNA géniques antisens. De même, chez la poule, DHCR24-DT a été identifié comme divergent du gène DHCR24 codant une enzyme clé dans la biosynthèse du cholestérol [108]. Cette classification est usuellement employée, mais présente un point faible important, elle dépend de l'annotation des PCG et de sa précision. Ainsi, un lncRNA considéré comme intergénique, peut se révéler en réalité être chevauchant en antisens par exemple après affinement du modèle PCG.

Dans la même mouvance, il est possible de classer les lncRNAs selon non pas uniquement les gènes codants, mais par rapport à des éléments régulateurs *cis*, cette classification permet de mettre en lumière l'idée de régions fonctionnelles du génome et de compléter celle présentée en amont. Bien que les lncRNA soient associés à la même classe, les notations correspondantes peuvent varier selon les différentes sources sans pour autant qu'une d'entre elles s'impose. De manière non exhaustive, on peut alors identifier :

- Les lncRNA associées à des régions *enhancer* distales (notés eRNA [139, 140], ncRNA-a [141, 142], ou encore elncRNA [143, 144]) stimulant généralement l'expression des gènes cibles tel que démontré avec ncRNA-a7 agissant sur le PCG SNAI1 [141].
- Les lncRNA associées à des régions promotrices actives en amont des TSS (notés pRNA [145], plncRNA [146], pancRNA [147], uaRNA [148], TSSa-RNA [149], PROMPT [150] ou encore PALR [105]).
- Les lncRNA associées à des régions génomiques ultra-conservées (notés TUCR pour T*ranscribed UltraConserved Regions*) à l'image du lncRNA Evf-2, situé dans une région intergénique bornée par les gènes Dlx-5 et Dlx-6 et identifié comme augmentant l'activité transcriptionnelle en influençant directement Dlx-2 [151]. Ces éléments sont en général très conservés chez des espèces proches telles que l'humain, la souris et le rat, mais également pour des espèces plus éloignées comme l'humain et la poule [152–154].

Un des points faibles des deux précédentes classifications est de considérer les lncRNA par rapport à une entité autre que ce soit un gène ou un élément régulateur, mais jamais comme élément existant en tant que tel. Ainsi, au lieu d'adopter une approche positionnelle, une des approches possibles est de considérer les lncRNA sur la base de leur structure tridimensionnelle. Cependant, si cela peut permettre d'apporter des informations sur leurs

fonctionnements, la modélisation et la validation des structures restent très complexes et couteuses et seule une poignée de lncRNA ont été caractérisés selon ces critères. Notons néanmoins que pour ceux ayant été caractérisés à ce jour, ils semblent adopter des structures modulaires et multidomaines, composées de plusieurs motifs structuraux distincts, mais interconnectés tels que déjà observés pour les PCG [155]. Ainsi, si les structures des lncRNA pourraient apporter des indications quant à leurs fonctions, il serait possible de classer ces lncRNA par leurs mécanismes.

De ce fait, à la lumière des données bio-informatiques et quelques fois expérimentales disponibles, des hypothèses fonctionnelles peuvent être émises. Cela peut ainsi guider les expériences à entreprendre pour les caractériser plus finement, mais cela peut s'avérer très complexe et très couteuse. Ainsi, à ce jour, les différents systèmes de classification des lncRNA se complètent pour décrire la grande diversité de ces ARN non-codants, notamment en lien avec leur localisation génomique et leurs structures. Si récemment, les bases de données comme Ensembl ou Refseq [156–158] tendent à diminuer le nombre de classes quitte à revenir à un unique groupe (« lncRNA »), ils intègrent ces différentes classifications en assignant différents biotypes tels que :

- « *3' overlapping ncRNA* » : Transcrit pour lequel les données expérimentales publiées confirment fortement l'existence d'un long (>200 bp) transcrit non codant qui chevauche le 3'UTR d'un locus codant pour une protéine sur le même brin.
- « *Antisense* » : Transcrit qui chevauche la partie génomique (c'est-à-dire exons ou introns) d'un locus codant pour une protéine sur le brin opposé.
- « *Macro lncRNA* » : lncRNA non épissé qui a une taille de plusieurs kilobases.
- « *Non coding* » : Transcrit dont on sait, d'après la littérature, qu'il ne code pas pour des protéines.
- « *Retained intron* » : Transcrit épissé alternativement, censé contenir une séquence intronique par rapport à d'autres transcrits codants du même gène.
- « *Sense intronic* » : Transcrit dans les introns d'un gène codant qui ne chevauche aucun exon.
- « *Sense overlapping* » : Transcrit qui contient un gène codant dans son intron sur le même brin.
- « *lincRNA (long intergenic ncRNA)* » : Long locus d'ARN non codant intergénique d'une longueur supérieure à 200 pb. Le transcrit ne possède pas de potentiel codant et peut ne pas être conservé entre les espèces.
- « *Bidirectional lncRNA* » : Transcrit provenant de la région promotrice d'un PCG, la transcription se déroulant dans la direction opposée sur l'autre brin.

Le but souhaité est à terme d'annoter finement les lncRNAs dans les bases de données génomiques, de la même manière que l'on classe précisément les gènes codants (oncogènes, suppresseurs de tumeurs, facteurs de transcription, récepteurs, kinases…). Cela facilitera ainsi l'identification des acteurs clés et des mécanismes de régulation impliqués dans les processus physiologiques et pathologiques, mais également permettra d'orienter les efforts de recherche tout en optimisant les coûts matériels et humains.

1.2.3. Les ARN longs non codants : des régulateurs aux modes d'action variés et aux conséquences encore peu identifiées

1.2.3.1. Interactions et mécanismes d'actions

Plusieurs travaux dans la littérature [102, 115, 159, 160] permettent d'avoir une vue d'ensemble du fonctionnement et des modes d'interaction des lncRNA. Cependant, dans le cadre du travail de la *review-1* (Figure 10) introduite précédemment, un état de l'art exhaustif a été dressé avec différents exemples appuyant chacun des mécanismes [114]. Cette introduction a ici été reprise en partie, le papier complet étant présenté en fin de §1.2.

Les lncRNA représentent une classe large et hétérogène de gènes régulateurs impliqués dans l'expression des gènes et peuvent agir à différents niveaux en utilisant divers mécanismes biologiques basés sur des interactions à tous niveaux, que ce soit avec l'ADN génomique, différents types d'ARN, les protéines et même une combinaison de ces composants.

Concernant l'ADN génomique, les lncRNA peuvent intervenir au niveau de :

- L'organisation nucléaire (*e.g.*, MALAT1 [161] / NEAT1 [162]) ;
- L'intégrité du génome en intervenant dans la stabilité des télomères du génome (*e.g.*, TERRA [163]) ou au niveau des marques d'histones pour le *silencing* (*e.g.*, Fendrr [164]) ou l'activation (*e.g.*, GATA3-AS1 [165]) de la transcription des gènes ;
- La formation de boucles pour connecter les *enhancers* aux régions promotrices (*e.g.*, MYMLR) [166].

Concernant l'ARN, les lncRNA peuvent intervenir au niveau de :

- L'épissage de l'ARN (*e.g.*, linc-HELLP [167]) ;
- La maturation des micros ARN (*e.g.*, CCAT2 [168] / uc.372 [169]) ;
- La stabilisation ou la dégradation de molécules comme les micros ARN (*e.g.*, SCR8 [170]), les ARNm (*e.g.*, PTB-AS [171] / TINCR [172]) ;
- Petits ARN en les hébergeant dans leurs introns [173] (*e.g.*, MCM7 [174] / DLEU2 [175]).

Concernant les protéines, les lncRNA peuvent intervenir au niveau de :

- La traduction des protéines (*e.g.*, BC1 [176] / MCM3AP-AS1 [177]) ;
- La modulation de l'activité protéique (*e.g.*, NORAD [178]) ;
- La stabilisation ou la dégradation des protéines (*e.g.*, PiHL [179] / MALAT1 [180]) ;
- L'hébergement de petits ORF qui codent pour des peptides (*e.g.*, CASIMO1 [181]/ DWORF [182]) ;
- La translocation de protéines du cytoplasme vers le noyau (*e.g.*, NRON [183]) ou du noyau vers le cytoplasme (*e.g.*, Discn [184]).

Indépendamment de cela, les lncRNA peuvent migrer vers d'autres cellules grâce aux exosomes (par exemple ZFAS1 [185] / GAS5 [186]).

Ainsi, grâce à leurs rôles clés dans la régulation de l'expression génique, les lncRNA sont par conséquent impliqués dans divers processus biologiques et physiopathologiques [134, 160, 187–189].

1.2.3.2. Vers une meilleure compréhension du rôle des ARN longs non codants dans les phénotypes d'intérêt

Même si la plupart des variations qui influencent les caractères quantitatifs se trouvent dans les régions non codantes (voir §3.1.3), qui présentent une multitude de lncRNA exprimés, l'effort de recherche s'est pour l'instant principalement concentré sur les mutations dans les gènes codants des protéines. En effet, ces PCG ont des conséquences plus fortes et plus faciles à repérer, et donc, à caractériser. Ainsi, en fonction de l'espèce considérée, le lien entre lncRNA et phénotypes n'est que peu établi. De plus, la plupart des études se concentrent sur l'expression spécifique des lncRNA dans des tissus ou des cellules spécifiques ou sur leur expression différentielle entre des groupes contrastés pour un trait, mais peu d'études proposent soit un mécanisme d'action, soit une démonstration *in vivo*, ou dans le cas idéal,

les deux. Il convient ainsi de différencier les études menées sur l'humain et la souris de celles menées chez les espèces domestiquées telles que présentées dans la *review-1* [103].

Par leurs fonctions régulatrices pléiotropes, les altérations des lncRNA sont associées à de nombreuses pathologies, particulièrement observées chez l'humain et la souris. De manière générale, des profils d'expression aberrants de lncRNA ont été observés dans tous les principaux types de cancer [190]. Ainsi, alors que certains lncRNA ont des activités oncogènes en favorisant la prolifération, la survie ou encore la migration des cellules tumorales, d'autre, à l'inverse, agissent comme suppresseurs de tumeur [102, 189, 190]. Au-delà du cancer, des dysfonctionnements de lncRNA sont également impliqués dans des maladies neurodégénératives telles que la sclérose latérale amyotrophique ou la maladie de Parkinson [191] ou encore dans des maladies cardiovasculaires [192] et auto-immunes [193].

En ce qui concerne les espèces animales, les phénotypes d'intérêt correspondent davantage à des caractères économiquement importants dans l'optique d'améliorer les productions, objectifs assumés des programmes de sélection artificielle mis en place au cours des dernières décennies même si ces objectifs ont évolués ces dernières années en lien avec des exigences économiques, sociétales et environnementales nouvelles [194]. De plus, les mécanismes biologiques et génétiques en lien avec ces phénotypes sont encore peu compris, alors que ces derniers pourraient offrir de nouvelles marges de progrès, que ce soit en termes de connaissances ou d'un point de vue économique. Dans cette mouvance, des centaines de milliers de signaux GWAS, principalement situés à l'extérieur des régions codantes, ont été identifiés afin de faire le lien entre génotype et phénotype d'intérêt. Cependant, comme évoqué précédemment, les régions non codantes, et notamment les lncRNA chez les espèces domestiquées incluant les espèces d'élevages, sont encore mal connues, bien plus que celles chez l'humain et la souris dont les connaissances restent elles-mêmes encore fragmentaires. À titre d'indication, comme montré dans la *review-2* (Figure 12) jusqu'à fin 2015, aucun lncRNA n'était décrit pour la poule et le bovin et seulement 135 pour le porc, alors que 14 896 et 6 830 lncRNA l'étaient chez l'homme et la souris, respectivement [103]. Cependant, pour tenter de palier à cette sous-connaissance du génome non codant, des projets internationaux tels que le consortium FAANG (*Functional Annotation of Animal Genomes*) ont vu le jour dans le but de cartographier les éléments fonctionnels et de mieux déchiffrer les relations entre le génotype et le phénotype [195, 196].

À titre d'exemple, les premiers lncRNA chez les trois principales espèces d'élevage ont été détectés au début des années 2010 dans la gonade mâle du porc [197], le muscle de la poule [198] et la peau de la vache. Depuis, le nombre de publications concernant les lncRNA dans ces trois espèces est en constante augmentation (une soixantaine est identifiée dans la *review-1* [103]) mais la plupart d'entre elles se concentrent encore une fois sur l'expression spécifique des lncRNA dans les tissus ou sur leur expression différentielle entre des races ou des groupes d'animaux contrastés. Soulignons par exemple, DHCR24-DT identifié comme associé au métabolisme lipidique en raison de son expression différentielle dans deux lignées de poules sélectionnées pour leur adiposité corporelle et divergentes pour de nombreux gènes en lien avec la synthèse du cholestérol (données personnelles). L'implication de lncRNA dans ce phénotype est également appuyée par sa localisation dans une orientation divergente par rapport au gène DHCR24 codant pour une enzyme clé de la synthèse du cholestérol à la fois chez la poule et l'homme, et de sa forte co-expression dans le foie hépatique avec ce gène dans plusieurs lignées de poules analysées à différents âges [108]. Des études sur les lncRNA ont par ailleurs été menées chez d'autres espèces animales, telles que la chèvre, le mouton, le lapin, le cheval, ainsi que chez d'autres espèces aviaires, telles que le canard ou l'oie [103].

L'étude des lncRNA ouvre donc des perspectives prometteuses pour mieux comprendre les bases moléculaires de nombreuses pathologies, pour identifier de nouveaux biomarqueurs ou dans le cadre de caractères économiquement intéressants. Actuellement, des approches expérimentales variées sont utilisées pour tenter de caractériser les mécanismes d'action des lncRNA, notamment des techniques de séquençage haut-débit, de cartographie épigénomique, d'imagerie cellulaire, et d'édition génomique comme le CRISPR-Cas9. Les progrès bio-informatiques permettent également d'intégrer ces données multi-omiques complexes pour modéliser les réseaux de régulation contrôlés par les lncRNA. L'exploration fonctionnelle des lncRNA reste néanmoins un grand défi et les conséquences phénotypiques peuvent être difficiles à mettre en évidence dans des conditions de laboratoire, du fait de leurs interactions transitoires avec de multiples partenaires, de leurs fortes dépendances à des conditions spécifiques et la dépendance à des modèles biologiques utilisés différents de ceux étudiés.

1.2.4. La nomenclature systématique des lncRNA : conventions et enjeux

La nomenclature des lncRNA suit des règles spécifiques, différentes de celles utilisées pour les gènes codant des protéines. En effet, les lncRNA étant dépourvus de fonction codante, leur nomination ne peut se baser sur la fonction de l'ARN encodé comme c'est le cas classiquement pour les gènes codants [199, 200]. Historiquement, les premiers lncRNA découverts ont reçu des noms évocateurs et descriptifs, reflétant la fonction supposée du transcrit. Par exemple, XIST qui signifie « *X inactive specific transcript* » et qui contrôle l'inactivation du chromosome X [135, 201]. De même, HOTAIR pour « *HOX Antisense Intergenic RNA* » qui régule l'expression des gènes HOX [202]. Cependant, à mesure que de nouveaux lncRNA ont été identifiés, notamment grâce aux techniques de séquençage haut-débit et face aux divers problèmes déjà rencontrés pour les gènes codants protéines (problèmes de doublons, de noms équivoques ou de manque d'homogénéité) [160], la communauté scientifique a ressenti le besoin d'établir une nomenclature systématique pour nommer les lncRNA de façon non ambigüe. Ainsi, le système d'attribution de nom des gènes chez l'Humain est porté par le *« Human Genome Organisatio*n » (HUGO) et plus particulièrement par un comité dédié, le « *HUGO Gene Nomenclature Committee* » (HGNC), qui publie et tient à jour des directives sur la manière de nommer les gènes ainsi qu'une base de donnée associée [200, 203]. Si ce consortium se concentre uniquement sur l'Humain, différents consortia ont vu le jour, que ce soit pour la souris (MGNC – *Mouse Gene Nomenclature Committee*) [204] ou le rat (RGNC – *Rat Genome Nomenclature Committee*) [205], d'autres vertébrés (VGNC – *Vertebrate Gene Nomenclature Committee* incluant 32 espèces en aout 2023) [206], mais également pour la poule (CGNC – *Chicken Gene Nomenclature Committee*) [207] ou encore le poisson-zèbre (ZNC – *Zebrafish Nomenclature Committee*) [208]. Cependant, malgré le poids du HGNC, la volonté de l'ensemble de ces groupes et de pouvoir translater les règles générales pour les différentes espèces [209] et ainsi de rester homogène pour pouvoir travailler sur un même gène dans différentes espèces (si ce dernier existe et est validé) aisément. Une règle indiquant que « La nomenclature ne doit pas contenir de référence à une espèce » vient par ailleurs confirmer cela [200]. Ainsi, la première mention de nomenclature pour les lncRNA datent de 2011 [210] avec des directives plus établies publiées en 2014 par Wright et al. [211]. Des propositions annexes seront faites [106] et cette nomenclature, en constante évolution, verra de nouvelles règles apparaitre en 2020 [212] et 2022 [213] notamment dans les publications de Seal et al.,

ou encore directement sur la page HGNC dédiée. Ainsi, les lncRNA se voient attribuer de préférence des symboles uniques sur la base de leur(s) fonction(s) publiée(s), à l'instar des gènes codant pour des protéines. Cependant face au peu de lncRNA caractérisés fonctionnellement (voir §1.2.3.2) mais au nombre important de lncRNA prédits (voir §1.2.1), les directives proposées se basent sur les classifications positionnelles évoquées précédemment (voir §1.2.2), et recommandent de nommer les lncRNA en suivant de manière séquentielle cette nomenclature :

N.B : Les nombres indiqués pour chaque catégorie correspondent aux lncRNA identifiés chez l'humain en août 2023.

- lncRNA hébergeant un miARN (micro RNA) ou snoARN (*small nucleolar ARN*) dans un exon ou un intron : nommé comme « hôte non codant » – [symbole miARN/snoARN]HG ; *e.g.,* MIR7-3HG/SNHG7; 97 [214] et 37 gènes [215] respectivement ;
- lncRNA divergeant par rapport à un PCG : nommé comme « divergent du PCG » – [PCG_{HGNC_name}]-DT ; *e.g.,* ABCF1-DT. ; 610 gènes [216] ;
- lncRNA contenu dans un intron d'un PCG : nommé comme «intronique du PCG » – [PCG_{HGNC_name}]-IT suivi d'un numéro unique ; *e.g.,* HAO2-IT1; 138 gènes [217] ;
- lncRNA chevauchant (sur un exon ou non) un PCG sur le même brin : nommé comme « chevauchant du PCG » – [PCG_{HGNC_name}]-OT suivi d'un numéro unique ; *e.g.,* PCBP2-OT1; 21 gènes [218]. Comme évoqué précédemment, ces cas représentent des cas limites, les lncRNA chevauchant les PCG sur leur partie exonique peuvent être considérés comme des transcrits alternatifs et ne représentent donc pas un locus à part entière.
- lncRNA chevauchant en antisens un PCG : nommé comme « antisens du PCG » – [PCG_{HGNC_name}]-AS suivi d'un numéro unique selon l'ordre de la soumission ; *e.g.,* ABCA9-AS1; 1932 gènes [219] ;
- lncRNA intergénique par rapport à des PCG : LINC suivi d'un numéro unique à cinq chiffres ; *e.g.,* LINC02998 ; 2347 gènes [220] ;

Chez l'humain, et par définition cela est/sera étendue aux autres espèces, si un lncRNA a un paralogue lncRNA dans le génome, ces derniers peuvent être nommés avec le symbole racine FAM (*family with sequence similarity*) : *e.g.*, FAM182A et FAM182B. ; 98 gènes [221].

La base de données permet donc d'accéder à l'identifiant HGNC et au symbole unique approuvé, mais également aux anciens symboles et alias connus de plus de 5 740 lncRNA [222] (475 non-systématique *vs.* 5 265 systématiques environ) facilitant les études sur les lncRNA et sur leur configuration avec les gènes environnants.

Ainsi, à la différence des gènes codants où la fonction protéique, si elle est connue, définit le nom, les lncRNA se voient essentiellement, pour l'instant, attribuer des identifiants non ambigus selon leur position génomique, en attendant d'être renommés de manière plus informative lorsque leurs fonctions seront connues. Cette stratégie de nomination systématique et normalisée facilite l'annotation et l'intégration des lncRNA dans les bases de données, même si des efforts sont encore nécessaire pour uniformiser les noms entre bases de données d'une même espèce. En effet, certaines n'adoptent pas la nomenclature évoquée ci-dessus ou ne l'ont intégré que récemment du fait de problème dans les *pipelines* pour considérer la correspondance entre deux modèles géniques (*e.g.*, NONCODE [112, 223] ; LNCipedia [224]). Le problème de nom se pose également entre plusieurs espèces lorsque la conservation du gène en question est effective. Par exemple, fin août 2023, chez la poule, seul un lncRNA (OIP5-AS1, aussi connu sous le nom de CYRANO) possède un nom suivant la nomenclature et avec une correspondance chez l'Homme. Les limites de cette nomenclature sont par ailleurs visibles avec ce simple exemple puisque, si ce lncRNA est bel est bien chevauchant de OIP5 chez l'humain, ce n'est pas le cas chez la poule où il serait classé comme divergent (DT).

Review-1

Mammalian Genome
https://doi.org/10.1007/s00335-021-09928-7

LncRNAs in domesticated animals: from dog to livestock species

Sandrine Lagarrigue[1] · Matthias Lorthiois[2] · Fabien Degalez[1] · David Gilot[3] · Thomas Derrien[2]

Received: 28 May 2021 / Accepted: 19 October 2021

Abstract
Animal genomes are pervasively transcribed into multiple RNA molecules, of which many will not be translated into proteins. One major component of this transcribed non-coding genome is the long non-coding RNAs (lncRNAs), which are defined as transcripts longer than 200 nucleotides with low coding-potential capabilities. Domestic animals constitute a unique resource for studying the genetic and epigenetic basis of phenotypic variations involving protein-coding and non-coding RNAs, such as lncRNAs. This review presents the current knowledge regarding transcriptome-based catalogues of lncRNAs in major domesticated animals (pets and livestock species), covering a broad phylogenetic scale (from dogs to chicken), and in comparison with human and mouse lncRNA catalogues. Furthermore, we describe different methods to extract known or discover novel lncRNAs and explore comparative genomics approaches to strengthen the annotation of lncRNAs. We then detail different strategies contributing to a better understanding of lncRNA functions, from genetic studies such as GWAS to molecular biology experiments and give some case examples in domestic animals. Finally, we discuss the limitations of current lncRNA annotations and suggest research directions to improve them and their functional characterisation.

Introduction

The last decade has witnessed the importance of the non-coding genome in the exhaustive characterization of genotype to phenotype relationships. Beside traditional protein-coding genes (mRNAs), animal genomes are pervasively transcribed into a myriad of short and long non-coding RNAs (Carninci 2005; Djebali et al. 2012; Mattick and Rinn 2015; Snyder et al. 2020) with various regulatory functions. Among these, long non-coding RNAs (lncRNAs) represent a vast and heterogeneous class of genetic elements with specific features in comparison with mRNAs. By definition, lncRNAs display very low coding-potential capabilities and are more tissue-specific and nuclear enriched than protein-coding genes (Cabili et al. 2011; Derrien et al. 2012). However, similar to mRNAs, they exert a variety of functions at either the transcriptional or posttranscriptional levels in cis or in trans (Ponting et al. 2009; Gil and Ulitsky 2019; Statello et al. 2021).

Given the interest for mapping to genomic regions the morphological, agronomical, or behavioural traits of domesticated animals, researchers have traditionally used genome-wide association studies (GWAS) to identify common polymorphisms associated with phenotypes of interest (Buniello et al. 2019). Yet, as in humans, many of the trait-associated variations identified by GWAS fall within non-coding intervals of the genome, reinforcing the need to deeply characterise the regulatory regions of domesticated species. Concomitantly, advances in high-throughput transcriptome sequencing technologies (RNAseq) has enabled the systematic exploration of this uncharacterised genomic space, first in human and model organisms (Djebali et al. 2012; Breschi et al. 2017) and more recently in other canonical and non-canonical organisms (Brown et al. 2014; Tagu et al. 2014). By combining RNAseq in numerous tissues or cell lines and at different developmental stages, it is now feasible to develop near comprehensive maps of coding and non-coding transcribed regions in order to refine the interpretation of genotype to phenotype studies in homogeneous populations of domesticated animals.

✉ Sandrine Lagarrigue
sandrine.lagarrigue@agrocampus-ouest.fr

✉ Thomas Derrien
thomas.derrien@univ-rennes1.fr

1 INRAE, INSTITUT AGRO, PEGASE UMR 1348, 35590 Saint-Gilles, France

2 Univ Rennes, CNRS, IGDR (Institut de Génétique et Développement de Rennes) - UMR 6290, 2 av Prof Leon Bernard, F-35000 Rennes, France

3 CLCC Eugène Marquis, INSERM, Université Rennes, UMR_S 1242, 35000 Rennes, France

Published online: 13 November 2021

Here, we review the current knowledge about lncRNAs mainly in dog, horse, cow, pig, and chicken chosen as main domesticated species and compare these lncRNA maps with respect to best-studied species in research such as human and mouse. The domestic dog (*Canis lupus familiaris*) is an exceptional case of species for tracking down genotype to phenotype relationships because pet dogs exhibit the most extreme phenotypic variations observed in terrestrial animals (Ostrander et al. 2017). This has been attributed to the particular history of dogs, from initial domestication events (> 14kya) of a now extinct grey wolf (*Canis lupus*) (Frantz et al. 2016) followed by intense breeding practices that led to the creation of modern purebred breeds during the Victorian era. However, this artificial selection for esthetical or behavioural traits has also led to the co-selection of morbid alleles that are now making dog breeds particularly predisposed to Mendelian diseases and cancers (Steenbeek et al. 2016). Dogs therefore represent an ideal genetic system to study phenotypically plastic traits and disease/cancer-related loci (Karlsson and Lindblad-Toh 2008). The pig, chicken and cow are livestock species which are the most used sources of animal protein worldwide, for the meat with 121, 114 and 67 Million tonnes produced worldwide and other products, e.g. eggs with 80 Million tonnes produced worldwide by laying hens (Food and Agriculture Organization of the United Nations 2021). These three species have been selected for multiple traits related to production (in terms of quantity and quality), efficiency, productive longevity, fertility, resilience, animal welfare and health. Among these three species, chicken has a particular status because of its phylogenetic characteristics, as birds and mammals diverged 300 mya. Finally, the horse (*Equus caballus*) is a key domesticated animal (~5 kya ago) from both cultural and economic aspects (Kalbfleisch et al. 2018) and has been selected for multiple traits (endurance, speed, appearance…).

For all these domesticated species, growing catalogues of long non-coding RNAs are being characterised, leading to increased examples of the association of lncRNAs with phenotypic traits of interest. However, lncRNA loci are still incomplete compared with protein-coding gene catalogues, partly due to the biological properties of lncRNAs. Therefore, only a handful of lncRNAs in domesticated animals have been associated with a probable causative effect or have been functionally validated. We thus emphasise the need to integrate complementary approaches for better annotating lncRNAs and for functionally validating trait-associated non-coding elements in the study of genotype to phenotype relationships.

Annotation of long non-coding RNAs in domesticated species

Transcriptome sequencing has revolutionized the process of genome annotation (Zhong Wang et al. 2009). RNAseq can be used to target different RNA populations of the cells, either with or without polyA tails. Except for a few studies mostly in human cells (Djebali et al. 2012), most of the annotated lncRNAs so far in pets and livestock species have been extracted from protocols employing polyA RNA selection. Once transcriptome sequences are available and quality-controlled, the bioinformatic process of annotating long non-coding RNAs basically involves three major steps (Table 1). The first one consists in *mapping* transcriptomic data (ESTs, cDNAs and now short and long RNAseq reads) onto a reference genome using a splice-aware mapper (e.g. STAR (Dobin et al. 2013)) in order to correctly model exon–intron junctions (Djebali et al. 2017). The second step aims at assemble mapped reads into known (already present in the reference annotation) and novel transcripts using dedicated *transcript reconstruction* tools [*e.g.* Cufflinks (Trapnell et al. 2010) or StringTie (Pertea et al. 2015)]. While the two first steps are common to both coding and non-coding genes, the third step focuses on classifying novel transcripts into mRNAs or lncRNAs by computing their *coding-potential* capabilities. An additional though optional step would involve the sub-classification of newly annotated lncRNAs with respect to the localisation and the direction of transcription of proximal mRNA transcripts in order to define lncRNAs classes such as lincRNAs (long intergenic ncRNAs) or antisense lncRNAs.

Based on dedicated annotation resources

LncRNA maps of domesticated species can be reached from several publicly available resources. As shown in Table 1, these resources use different computational tools at each main step of the RNAseq processing pipeline described above (Table 1). Furthermore, the total number of lncRNA genes and transcripts vary substantially between domesticated species and do not currently scale with the number of lncRNA in human and mouse catalogues (Table 1).

One of the most widely used resources for extracting gene annotations is provided by the Ensembl genome browser (Aken et al. 2016; Howe et al. 2021). Ensembl provides genome-wide annotations of protein-coding and non-coding RNAs for more than 250 vertebrates, including many domesticated animals. In human or canonical model organisms (*e.g.* mouse), the specific process of annotating

Table 1 Bioinformatic tools for annotating and classifying lncRNAs from multi-species databases

Database name	Read mapping	Gene modelling	Coding-potential assessement	Number of lncRNA genes/transcripts by species *(genome assembly version)*						
				Human	Mouse	Cow	Pig	Chicken	Dog	Horse
Ensembl (v104)	BWA	Exone-rate	ORF and PFAM alignment[a]	16 896/46 960 (GRC g38. p13)	9972/12 601 (GRC m39)	1488/2199 (ARS_UCD1.2)	6979/9367 (Sscrofa 11.1)	5506/8870 (GRC g6a)	7083/12 283 (Can Fam 3.1)	7244/11 978 (Equ Cab 3.0)
NCBI (v105)	Minimap2 (long read) Spilign (short-read)	Gnomon	Gnomon	16 375/27 838 (GRC g38. p13)	13 317/23 542 (GRC m39)	5183/7254 (ARS_UCD1.2)	5605/9292 (Sscrofa 11.1)	5147/8233 (GRC g6a)	10 823/19 248 (Can Fam 3.1)	6789/10 850 (Equ Cab 3.0)
NONCODE (v6.0)	Literature parsing with RNAseq key words + "CuffCompare" to deal with overlapping features		Comparison with RefSeq + "CNIT"	96 411/173 112 (GRC g38)	87 890/131 974 (GRC m39)	22 127/23 515 (UMD 3.1)	17 811/29 858 (Sscrofa 10.2)	9527/12 850 (galgal4)	NA	NA

The number of lncRNAs found for each species with the corresponding assembly used is also presented

[a]Gene models containing a substantial open reading frame (ORF) and protein domains (e.g. from Pfam) are classified as coding. For human and mouse annotations, additional manual curations from Gencode

long non-coding RNAs combines automated annotation from RNAseq data processed by the Ensembl gene build pipeline and manual curation by the HAVANA/Gencode group (Frankish et al. 2019). The Gencode database (version 37), which is synchronised with Ensembl, has compiled 17 948 human lncRNA genes (~48 000 transcripts) and 13,186 mouse lncRNA genes (~18 000 transcripts) (version M26). For other species, including domesticated animals, the description of the built lncRNA catalogues has been less detailed to date and does not include manual curation which most likely impacts the quality of these annotations. In addition, in contrast to human and mouse Ensembl catalogues, only intergenic genes (lincRNAs) are referenced, meaning that other biotypes such as antisense exonic or sense intronic transcripts, are not reported for domesticated species.

The number of Ensembl lncRNA genes varies greatly between the 5 major domesticated species. For instance, 1480 lncRNA genes have been identified in the cow and approximately 7000 in the horse, dog, and pig, whereas the number of protein-coding genes (mRNAs) remains more stable (~20,000) (Fig. 1A). Similar to mouse, the number of lncRNA transcripts/isoforms per gene in the cow and dog ranges from 1.4 to 1.7 lncRNA transcripts per gene, respectively, which is significantly lower than the 1.8 and 2.5 mRNA transcripts per gene for protein-coding genes in the respective species. This might be due to the difficulty to identify lowly expressed lncRNA isoforms by RNAseq methodologies (Fig. 1A). When comparing the length of lncRNA transcript sequences across domesticated species (Fig. 1B), one could note that pig and chicken lncRNA transcripts are significantly longer than those in other mammal species (Mann–Whitney *U* tests, p values < 2.2e-16). Interestingly, the recent annotations of the new *sus scrofa* and *gallus* assemblies have benefited from the use of long-read RNAseq (LR-RNAseq) (PacBio Iso-Seq from nine adult porcine tissues (Warr et al. 2020; Beiki et al. 2019) and from originally two and now six addition chicken tissues (Kuo et al. 2017; https://www.ensembl.org/Gallus_gallus/Info/Annotation), which might have enabled global extensions of transcript models as this trend has also been observed for protein-coding genes (Fig. 1B).

As every automatic modelling process, the Ensembl gene build pipeline might also suffer from incorrect annotations. A closer inspection of the Ensembl-based catalogues of lncRNAs in the five domesticated species identified the probable misclassification of some mRNAs as long non-coding transcripts. For instance, between 5.5% lncRNAs in horse and 11.8% lncRNAs in cow were classified as protein-coding by the FEELnc program (Wucher et al. 2017). When searching for the longest ORFs, either partial (*i.e.* missing start codon) or full (having both a start and stop codons), in these "ambiguous" transcripts (Fig. 1C), the ORF appears

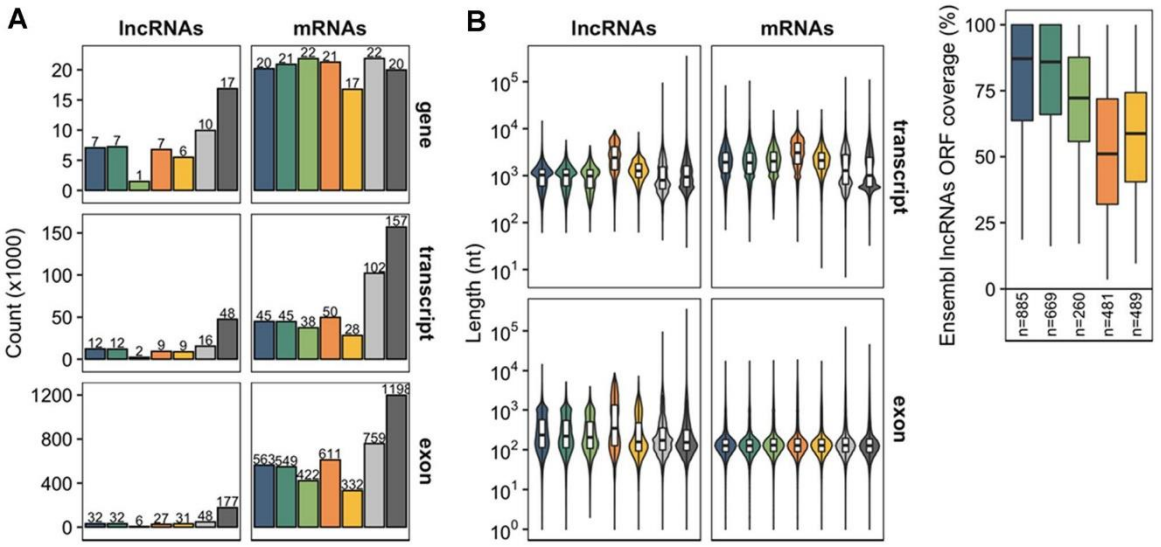


Fig. 1 Characterization of lncRNA and mRNA gene structures in 5 domesticated animals (dog, horse, cow, pig, and chicken, respectively in dark green, orange, purple, pink, and light green) in comparison with mouse and human annotations (light and dark grey respectively) extracted from Ensembl (v103). **A** Comparison of the number of lncRNA and mRNA genes, transcripts, and exons (number of lncRNA and mRNA features are indicated on top of each bar). **B** Boxplot distributions of the length of lncRNA and mRNA transcripts and exons. **C** ORF coverage of Ensembl-based lncRNAs annotated as protein-coding by the FEELnc program

to cover a large fraction of the annotated RNA sequences (median = 51% in pig to 82% in dog) despite the fact that it should have been filtered (Aken et al. 2016) (Table 1). Therefore, a high ORF coverage would suggest that these transcripts might represent bona fide protein-coding transcripts and exclude the possibility that they correspond to lncRNAs harbouring small ORFs (smORFs) (Bazzini et al. 2014; Ruiz-Orera et al. 2014).

Despite these shortcomings, the Ensembl resource is extremely useful for the scientific community working on non-model organisms because it provides a versioned, stringent, and freely available set of gene/transcript structures (both coding and non-coding) at the basis of most downstream bioinformatic analyses.

Besides Ensembl, several more recent databases also provide extensive annotations of non-coding genes based on different computation pipelines (Table 1). For instance, the NONCODE database (Zhao et al. 2016) is specifically dedicated to the annotation and bioinformatic characterization of long non-coding RNAs in animals and plants. The integration of lncRNAs in NONCODE makes use of the CuffCompare tool from Cufflinks (Trapnell et al. 2010) in order to combine and filter multiple sources of lncRNA annotations. One advantage of NONCODE over Ensembl is that it involves the use of a published coding-potential assessment tool, CNIT for Coding-Non-Coding Identifying Tool (Guo et al. 2019), an updated version of the CNCI program (Sun et al. 2013), to discriminate reconstructed coding from non-coding gene models. One limitation though is that NONCODE only includes lncRNA catalogues for 16 animal species, excluding dog and horse for instance. Whereas, in the case of Ensembl-matched species, the number of lncRNA transcripts is significantly higher with 9527, 17,811, and 22,227 lncRNA loci for chicken, pig, and cow, respectively. In addition, NONCODE provides a detailed characterization of annotated lncRNAs based on phylogenetic conservation, disease association, as well as lncRNAs overlapping SNPs/GWAS hits. Historically, the first specific database of lncRNAs dedicated to livestock species was the domestic-animal lncRNAs database (ALDB) (Li et al. 2015), although this database seems not to have been updated since 2016. Using a rather out-dated bioinformatic pipeline including the TopHat mapper and the CPC tool for assessing coding-potential, ALDB comprises 6151 (8923), 7381 (12 103), and 5213 (8250) lincRNA loci (transcripts) for chicken, pig, and cow, respectively. Finally, it is also worth mentioning the NCBI reference sequence database (RefSeq) that provides automatic annotation of lncRNAs and mRNAs in > 55,000 organisms, including domesticated species. In particular, NCBI/RefSeq makes use of the "eukaryotic genome annotation pipeline" with the Gnomon program, which combines homology searching with ab initio modelling (O'Leary et al. 2016) and comprises 10,823,

5 147, 5 605, and 5 183 lncRNA loci in dog, chicken, pig, and cow, respectively (Table 1).

Although these publicly available catalogues represent a rich resource for digging into trait-associated loci, involving annotated lncRNAs, a limited genomic overlap still exist between these annotations (Fig. 2), most likely reflecting the high specificity of lncRNA expression profiles and the different origins of the input transcriptomic sequencing data.

De novo transcriptome reconstruction of new long non-coding RNAs

The democratization of RNAseq combined with efficient bioinformatic tools to rapidly process transcriptome data have allowed researchers working on domesticated species to build their own catalogues of lncRNAs.

Long non-coding RNA studies and atlas in dogs

The scientific community provided a first dog reference genome assembly, together with an annotation of ~ 20,000 protein-coding genes, of a boxer breed in 2005, making the dog the fifth mammal to be sequenced (Lindblad-Toh et al. 2005). However, a comprehensive catalogue of coding and non-coding/regulatory elements for the interpretation of the many GWAS signals lying outside of annotated mRNAs and for the eventual identification of the actual causal mutations was not provided until 2014. At that time, Hoeppner and colleagues combined RNAseq data from 10 distinct canine tissues to build ~ 7200 lincRNA transcripts and 4600 antisense lncRNAs (Hoeppner et al. 2014). In 2017, thanks to the collection of novel canine RNA samples provided within the framework of the European LUPA consortium (Lequarré et al. 2011), Wucher et al. integrated 20 additional RNAseq data to build a new canine reference annotation (Wucher et al. 2017). Using the dedicated FEELnc program to automate the annotation of lncRNAs and their genomic classification (lincRNA, antisense, and other subclasses), the authors provided an extended set of canine lncRNAs comprising 22,880 lncRNA transcripts gathered into 10,444 gene loci. A deeper analysis of this extended RNAseq dataset revealed that, as in humans, canine lncRNAs are more tissue-specific than protein-coding genes (44 versus 17%, respectively) with 65% of all tissue-specific lncRNAs expressed in canine testis (Le Béguec et al. 2018). This catalogue was

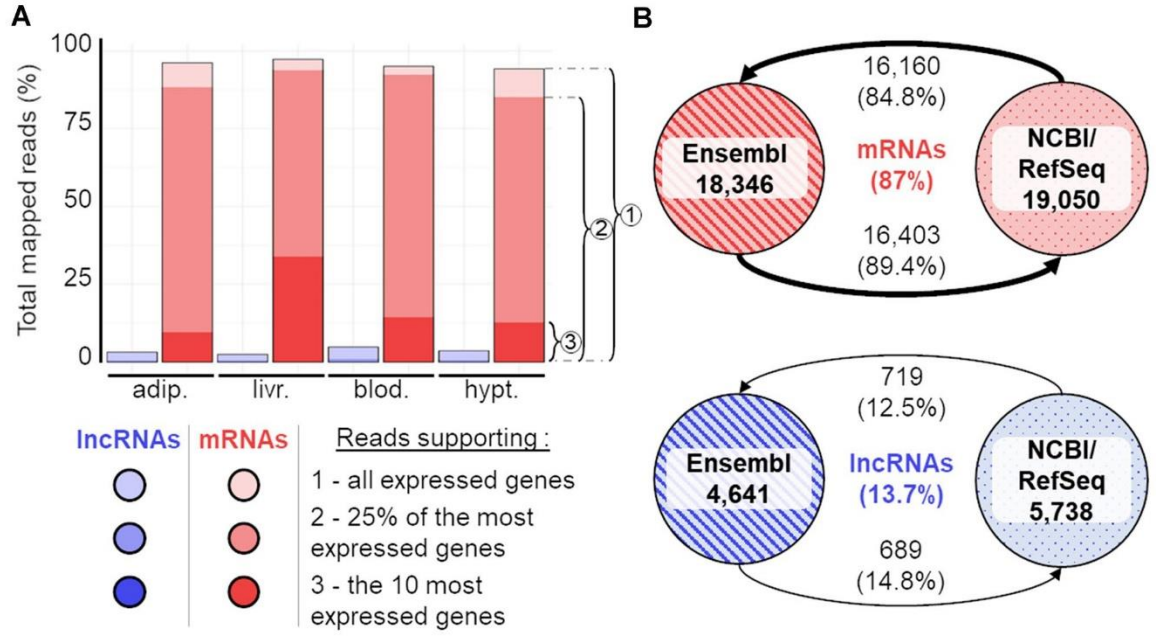


Fig. 2 Distribution of reads supporting lncRNAs and mRNAs (**A**) and gene overlap between NCBI and Ensembl resources according to both biotypes (**B**). **A** For each gene biotype (lncRNAs in blue and mRNAs in red), the dark, intermediate and light shades correspond to the percentage of reads supporting all expressed genes, 25% of the most expressed genes and the 10 most expressed genes respectively. RNAseq data correspond to the chicken PRJEB28745 project and 4 tissues (*adip* adipose tissue, *livr* liver, *blod* blood, *hypt* hypothalamus) of the same population (Rhode Island Red). B) Percentages of chicken lncRNA gene overlap—using 1 bp or more—between the GRCg6a—V104 Ensembl and NCBI gene catalogues. Note that these overlaps have been computed at the gene level given the uncertainty of isoform modelling with short-reads as explained in the main text

first analysed in the context of dog breed phenotypic variations, such as the "drop ear" phenotype, in which case using GWAS one lncRNA was found to be closely associated to the *MSRB3* gene involved in human deafness (Plassais et al. 2019) (see below "GWAS hits involving lncRNAs"). Furthermore, given the combined interest for lncRNAs as potential cancer drivers/biomarkers (Huarte 2015; Vancura et al. 2021) and dogs as natural and thus immunocompetent models for cancer analyses (Prouteau and André 2019), canine lncRNAs were analysed in three canine breeds (poodles, Labradors, and golden retrievers) predisposed to mucosal melanomas (MM). Using RNAseq in tumour and adjacent matched control tissues, more than 400 lncRNAs were shown to be differentially expressed between healthy and diseased animals, with 26 of these lncRNAs being reported to be conserved in humans (Hitte et al. 2019). In addition, while MM is a rare cancer in humans, the high frequency of MM in particular breeds enabled the identification of ~ 10 breed-specific lncRNAs, which were shown to be specifically differentially expressed in one breed versus the others (Hitte et al. 2019). Beside melanomas, a number of studies have established lncRNA atlases in canine cancers, such as B-cell lymphoma (DLBCL) (Cascione et al. 2019; Verma et al. 2015) or canine kidney cancer (MDCK) (Qiao et al. 2020) (Table 2A) and also linked GWAS hits to overlapping lncRNAs such as in hematopoietic cancers (Hédan et al. 2021).

Long non-coding RNA studies in farm animals

Concerning livestock species, artificial selection programs, including recent genomic selection methods, have led to spectacular gains in economically important traits over the last decades (Hill 2016). However, there is little understanding of the biological mechanisms underlying such phenotypes, the knowledge of which could offer new margins of progress, such as, making genomic selection methods more robust or better exploiting the genotype-environment interactions. Therefore, a new goal of the scientific community in the animal genetic field is to provide the functional annotation of the genomes of farm animals to elucidate the hundreds of thousands of GWAS signals [160 659, 31 455, and 12 783 in the three major livestock species of cow, pig, and chicken, respectively (Hu et al. 2019)], which are known to be mainly located outside the ~ 20,000 coding regions. Chicken was the first species with a large genome to be sequenced in 2004, just after those of human and mouse (International Chicken Genome Sequencing Consortium 2004). However, the knowledge of non-coding regions in farm animals has not kept up with that in humans. Until December 2015 (Ensembl version 83), no lncRNAs were described for chicken and cow and only 135 were reported for pig, as contrasted with 14,896 and 6830 lncRNAs reported in human and mouse, respectively. This poor knowledge of the non-coding genome annotation has led to a coordinated international action to accelerate genome to phenome, termed the Functional Annotation of Animal Genomes (FAANG) project, whose aim was to produce comprehensive maps of functional elements in the genomes of livestock species to better decipher the genotype to phenotype relationships (Andersson et al. 2015). As part of FAANG, two studies have recently provided a multispecies lncRNA annotation using 8 tissues of 2 biological replicates of 3 species, namely chicken, pig, and cattle (Kern et al., 2018) and 3 tissues of 4 biological replicates of 4 species, namely chicken, pig, goat, and cow (Foissac et al. 2019).

The first lncRNAs in the three major livestock species were detected in the male gonad (Esteve-Codina et al. 2011), muscle (Li et al. 2012), and skin of the pig, chicken and cow, respectively, in the early 2010s. Since 2015, the number of publications regarding these three species has been constantly growing, with most of them focusing on the tissue-specific expression of lncRNAs or their differential expression between breeds or animal groups contrasted for an economically important trait in the species of interest (Table 2B). LncRNA studies have also been conducted in other livestock species, such as goat, sheep, rabbit, horse, as well as in other avian species, such as duck or geese (Table 2C). However, to our knowledge no studies have been performed in turkey and quail despite the identification of 1038 and 5090 lncRNAs in these two species, respectively, in the latest Ensembl annotation version (v104).

In most of these studies, a few lncRNAs have been highlighted from the lncRNA catalogues as associated to the trait of interest because of their significant differential expression between two animal groups of interest and their co-expression with a close protein-coding gene that can be used as a proxy to infer possible functions for the lncRNA, especially when the lncRNA is conserved in multiple species. For instance, the *linc-SABT1* (that should be renamed to *SATB1_DT*) has been associated with resistance to Marek's disease (MD), because of (i) its high expression in infected birds of the Marek's disease resistant line, and (ii) its location in the divergent orientation of the *SATB1* gene known to regulate chromatin structure and control a large number of immunity genes (He et al. 2015). The *DHCR24-DT* has been associated with lipid metabolism because of (i) its differential expression in 2 divergent lines selected for body adiposity, (ii) its location in a divergent orientation of the *DHCR24* gene coding for a key enzyme of the cholesterol synthesis in chicken and human, and (iii) its high hepatic co-expression with this mRNA gene in several chicken lines (layers and broilers) analysed at different ages (young and adult stage) (Muret et al. 2017).

Table 2 LncRNA studies associated with trait-related tissues in dog and livestock species

Tissues	Related traits/disease	Species	References
A. Dog			
Retina	X-linked progressive retinal atrophy	Dog	(Appelbaum et al. 2020)
Various	Breed morphology (*e.g.* "drop ear")	Dog	(Plassais et al. 2019)
Mucosal and skin tissues	Mucosal melanoma	Dog	(Hitte et al. 2019)
Lymph node	Lymphoma	Dog	(Verma et al. 2015; Cascione et al. 2019)
B. Three major species: pig, chicken, and cow*			
Muscle	Growth performance and meat quality	Pig	(J. Sun et al. 2017; Zou et al. 2017a, b; Zou et al. 2017a, b; Li et al. 2020)
		Chicken	(Li et al. 2012; Li et al. 2019; Ren et al. 2018a, b; Cai et al. 2017)
		Cow	(Choi et al. 2019; Li et al. 2020)
Mammary gland	Milk production and quality	Cow	(Tong et al. 2017; Yang et al. 2018; Ibeagha-Awemu et al. 2018; Zeng et al. 2019)
Immunity tissues	Disease or resistance against pathogenic infections	Pig	(Fang et al. 2019)
		Chicken	(Qiu et al. 2017; Hu et al. 2018; Ren et al. 2018a, b; You et al. 2019; Dai et al. 2019; Li et al. 2021; Zhang et al. 2021)
		Cow	(Özdemir and Altun 2020)
Male sexual organs	Male reproduction traits	Pig	(Esteve-Codina et al. 2011)
		Chicken	(Liu et al. 2017a, b; Zou et al. 2020)
		Cow	(Wang et al. 2019a, b; Gao et al. 2019)
Female sexual organs	Female reproduction traits	Pig	(Wang et al. 2016; Wang et al. 2019a, b)
		Chicken	(Liu et al. 2018; Adetula et al. 2018; Peng et al. 2019; Yin et al. 2020; Zou et al. 2020)
Liver and adipose tissues	Body lipid reserves and metabolic efficiency	Pig	(Wang et al. 2017; Miao et al. 2018; Kumar et al. 2019)
		Chicken	(Muret et al. 2017; Zhang, et al. 2017; Zhang 2017a, 2017b; Wu et al. 2018; Xu et al. 2019; Muret et al. 2019; Chen et al. 2019; Ning et al. 2020)
		Cow	(Nolte et al. 2019; Kong et al. 2020; Alexandre et al. 2020)
Intestine	NA	Cow	(Weikard et al. 2018; Nolte et al. 2019)
Spleen	NA	Pig	(Che et al. 2018)
		Chicken	(You et al. 2019)
C. Other livestock species*			
- Liver and cerebral parietal lobe - Placenta - Eight tissues		Horse	(Dahlgren et al. 2020; Pu et al. 2020; Scott et al. 2017)
- Skin - Endometrium - Ovary and follicle		Goat	(Ren et al. 2016; Hong et al. 2020; Lian et al. 2020; Zhao et al. 2020)
- Multiple tissues - Wool - Pituitary - Oocyte development - Consensus set of ruminant lncRNAs		Sheep	(Bakhtiarizadeh et al. 2016; Yue et al. 2015; Zheng et al. 2019; Yang et al. 2020; Wang et al. 2020; Bush et al. 2018) consensus set of ruminant lncRNAs provided by Bush et al. 2018
- Muscle - Adipose tissue - Skin - Embryos		Rabbit	(Kuang et al. 2018; Wang et al. 2018; Zhao et al. 2019; Ding et al. 2021; Kuang et al. 2020)
- Ovary - Brain, lung and spleen - Embryo fibroblast cells		Duck	(Ren et al. 2017a, 2017b; Lu et al. 2019; Y. Lin et al. 2020a,b)

for Ensembl and 129 samples from different projects for NCBI/RefSeq) in comparison to the thousands of RNAseq samples generated over the past decade and publicly available in ENA or SRA databases. Therefore, these reference gene sets do not recapitulate the diversity of tissues, ages and physiological stages of lncRNA expression patterns. Consequently, lncRNA gene models are highly sample-dependent in comparison to more broadly expressed mRNAs, as illustrated by the little overlap of lncRNA loci between Ensembl and NCBI/RefSeq (about 13.7%), whereas almost all mRNA loci are common to both resources (87%) (Fig. 2B).

Finally, as previously illustrated in Table 1, lncRNA databases also make use of different bioinformatic tools at each step of the lncRNA annotation process (Table 1). This most likely influences gene structure boundaries (especially given the limitations of tools for the reconstruction of full transcripts from short-read RNAseq) together with the correct attribution of gene biotypes (mRNA versus lncRNA), and therefore, the extent of overlap between lncRNA sets.

In conclusion of this section, unlike protein-coding genes, genome annotation for lncRNAs (transcript and gene loci) requires considering the entire diversity of tissues, stages, conditions available in public sequences databases. In combination with standard computational procedures and benchmarked tools, the inclusion of many more projects and associated RNAseq samples within the same species both using short-read RNASeq and, in the coming years, long-read RNAseq technologies will most likely increase the completeness of lncRNA sets in domesticated animals.

Long non-coding RNAs and comparative genomics

Comparative genomics, defined as the comparative study of the structure and function of the genomes of different species, is a common method to identify new genes and their functions, and thus to more accurately annotate new genomes (König et al. 2018). However, although the approaches used for protein-coding genes are quite efficient, they have been revisited for the long non-coding genes (lncRNAs) due to their structural and functional specificities.

Over the past decade and linked to the growing interest for lncRNAs, multiple studies have used comparative genomic approaches to detect and annotate novel lncRNAs across phylogenetically divergent species. (Necsulea et al. 2014; Hezroni et al. 2015; Sarropoulos et al. 2019). However, a set of annotated genomes and a bioinformatic method to compute the distance/similarities between the source and target genomes are required. So, even though the catalogues of lncRNAs in many species have been increasing, especially due to the standardization of RNAseq-based methods, lncRNA repertoires of domesticated species remain mostly incomplete, as underlined before. If the incomplete annotation of lncRNAs represents one of the issues for the comparative study of conserved lncRNAs, the phylogenetic divergence between targeted species is also an important parameter to be considered.

Indeed, lncRNAs evolve very fast and, usually, the higher the evolutionary distance between two species, the fewer the number of orthologous lncRNAs (Bu et al. 2015; Chen et al. 2016; Hezroni et al. 2015; Kern et al. 2018; Washietl et al. 2014; Necsulea et al. 2014). Moreover, the rates of birth and death of lncRNAs seem to be very high, even in closely related species, as shown by Kutter et al. in rat and mouse species, where half of the intergenic lncRNA loci have been gained or lost since the last common ancestor (20 My) (Kutter et al. 2012). And so some lncRNAs might appear as derived from a lost protein-coding gene (Duret 2006; Hezroni et al. 2017). Finally, even if the genomic sequence of a lncRNA is conserved, its expression profile in matched tissues might differ between species (comparative transcriptomics) (Washietl et al. 2014).

In the case of the domesticated species, these evolutionary distances are quite heterogeneous (Fig. 3A). Indeed, even though most of the species of the "domesticated" group diverged from human ~ 96 mya, the evolutionary distances within the group are very variable. For example, the closest species are "goat" and "cow" that share a common ancestor around 25 mya., whereas "pig" diverged 62 mya. The chicken appears as an outlier because it diverged 300 mya. Interestingly, some lncRNAs appear to be conserved over a large time-scale possibly due to their common function in all eukaryotes (Kern et al. 2018; Wiberg et al. 2015).

Based on all these observations and considering the availability of adequately annotated genomes, several-related approaches have been used to perform comparative genomic analyses of lncRNAs. The first one, which was usually used for protein-coding genes, is based on the alignment of the primary sequences of genes on the target genome. However, although this technique works relatively well for mRNAs, it needs to be adapted for lncRNAs. Overall, around 70% of lncRNAs have no sequence orthologues (e.g. given a certain threshold of sequence similarity and alignment length) in species that have diverged for over 50 mya (Hezroni et al. 2015). Furthermore, not all parts of a lncRNA sequence evolve at the same rate. LncRNA exons are more stable than intergenic sequences and mRNA introns (Cabili et al. 2011). So, only a few "patches" of sequences (*e.g.* short conservation islands), potentially corresponding to RNA or protein binding regions, seem to be conserved and are generally located in lncRNA exons and promoters (Noviello et al. 2018; Darbellay and Necsulea 2020). These patches are significantly shorter than those located in mRNAs, are found in only one or two exons, and can tolerate large rearrangements. Quinn et al. considered that only 10% of the

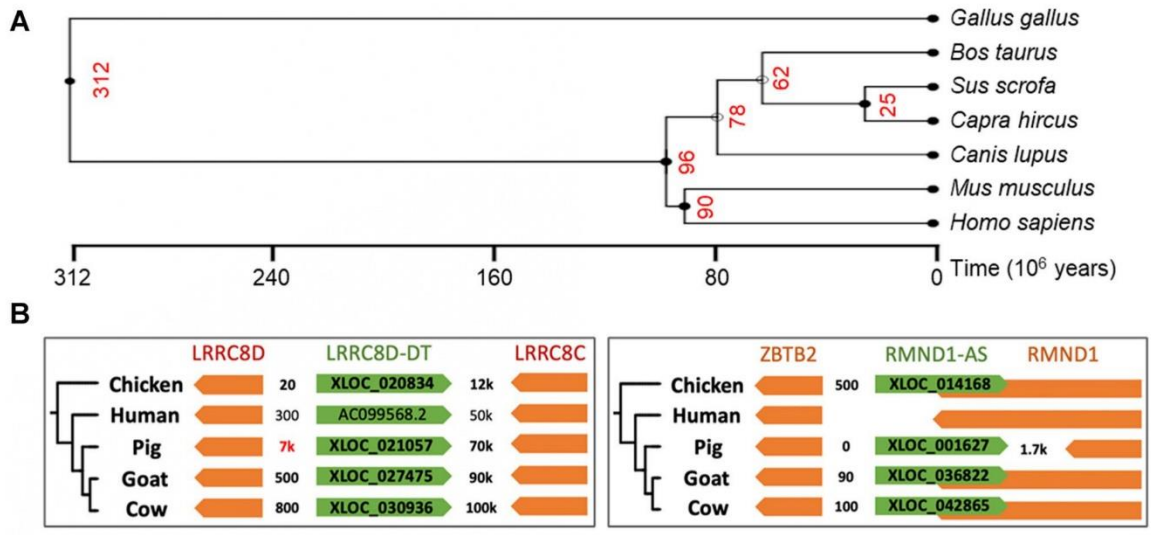


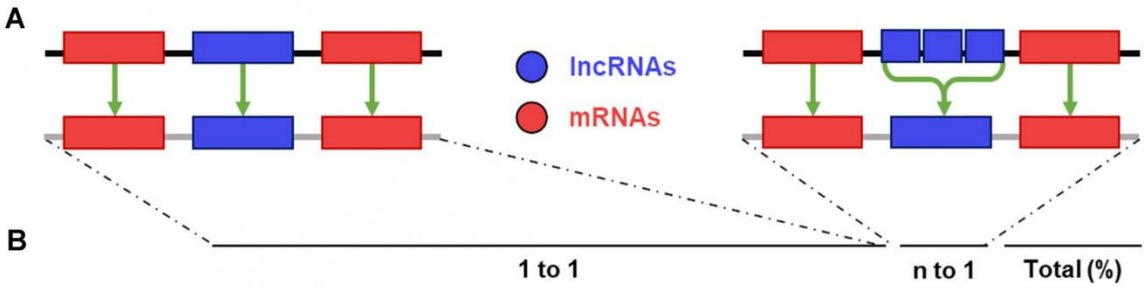
Fig. 3 Phylogenetic divergence between domesticated species, mouse, and human. **A** Red numbers correspond to the common ancestor of different species. This tree was generated using the TimeTree database (Kumar et al. 2017). Distances were calculated from estimated molecular time. **B.** Genomic conservation of 2 lncRNAs (in green) in divergent position extracted from Foissac et al. (Foissac et al. 2019)

sequence might be sufficient to support the function of a lncRNA (Quinn and Chang 2016). Recently, a new tool, called lncLOOM, based on a graph representation of a multiple sequence alignment (MSA) and integer linear programming, has been published for the functional prediction of lncRNA short motifs positionally conserved between species (Ross et al. 2021). Applied to vertebrate species, the tool allowed the identification of functional domains in known lncRNAs, such as *Cyrano* and *CHASERR,* as well as in the 3′-UTR of protein-coding transcripts (Ross et al. 2021).

However, while lncRNA gene structures change rapidly and might therefore be an obstacle to the detection of homologous sequences, other important features can be used in the detection of lncRNAs by comparative genomics. Indeed, lncRNAs are more tissue specific than protein-coding genes, which can help refine predicted functions (Guttman et al. 2011). Such a characteristic shows the importance of working with matched tissue(s) between species in the case of comparative transcriptomic approaches. Interestingly, the oldest conserved lncRNAs are generally expressed in tissues related to embryonic development (Necsulea et al. 2014; Washietl et al. 2014). Another major attribute of the biology of lncRNAs is related to their positional conservation (synteny) between species genomes. This trend has been observed between human and mouse, as well as in the case of comparative genomic analysis of domesticated animals (Foissac et al. 2019) (Fig. 3B). A possible explanation could be their potential function related to gene regulation through the reorganization of local chromatin structure. To identify such positionally conserved lncRNAs, the identification of positionally conserved neighbour genes, usually mRNAs, is initially required; if these genes are orthologous in the targeted species, they will also define a conserved syntenic interval for lncRNAs. Using this strategy, a few studies have found positionally conserved lncRNAs within distant species (Hezroni et al. 2017, 2015; Sarropoulos et al 2019; Muret et al. 2017, 2019; Jehl et al. 2020).

Using a similar approach, we estimated the number of syntenic lncRNAs among seven species including domesticated species (except horse), mouse and human (Fig. 4B). As depicted in Fig. 4A, we have searched for lncRNAs corresponding to strict one-to-one equivalences (termed "1–1") for all the species-pairs. In a second step, we considered the "n–one" orthologous lncRNAs ("n–1") defined as n adjacent lncRNA loci in one of the six species related to a single syntenic lncRNA in the human species which is considered here to be the species with the most accurate annotation of lncRNAs.

As expected, the smaller the phylogenetic distance between species, the higher the number of orthologous lncRNAs. For instance, we observed with the human species between 190 and 628 orthologous "1–1" lncRNAs for the chicken and mouse species, respectively. For the other livestock species, between 119 Ensembl lncRNAs in cow and 282 lncRNAs in pig can be considered as syntenically conserved with a human lncRNA using the strict definition "1–1". It is important to note that the comprehensiveness of a species-specific lncRNA catalogue has a major impact on



		1 to 1							n to 1	Total (%)
		Chicken *	Chicken	Dog	Pig	Goat	Cow	Mouse	Human	Human
		24,835	*5,504*	*7,077*	*6,790*	*2,705*	*1,480*	*8,948*	*16,887*	*16,887*
Chicken *	*24,835*	-	0	130	155	84	76	333	427	2150	2577 *(10.4)*
Chicken	*5,504*		-	104	111	85	68	196	190	256	446 *(8.1)*
Dog	*7,077*			-	161	95	76	233	266	362	628 *(8.9)*
Pig	*6,790*				-	108	95	264	282	338	620 *(9.1)*
Goat	*2,705*					-	85	149	144	81	225 *(8.3)*
Cow	*1,480*						-	140	119	42	161 *(10.9)*
Mouse	*8,948*							-	628	588	1216 *(13.6)*
Human	*16,887*								-	-	-

Fig. 4 Syntenic conservation of lncRNAs across 7 species. **A** schema of "1–1" and "n–1" principles of positionally conserved lncRNAs. The "1–1" corresponds to the case of a strict and unique syntenic equivalent in both species located in-between two adjacent "1–1" protein-coding genes. The "n–1" corresponds to the case of multiple lncRNA loci in the analysed species that corresponds to an unique lncRNA in human located between the two "1–1" protein-coding genes. **B** Number of lncRNA for each homology category across species with numbers of lncRNA loci (in italic) extracted from Ensembl (v104). The "*" indicates the chicken lncRNA-enriched annotation anchored on the v101 (equivalent to v104) Ensembl resource (Jehl et al. 2020)

the number of annotated orthologous lncRNAs. Indeed, we observed 427 (2150) versus 190 (256) "1–1" ("n–1") orthologous lncRNAs between chicken and human species when comparing the two lncRNAs chicken atlas (v104 Ensembl catalogue versus lncRNA-enriched v104 Ensembl catalogue from Jehl et al. 2020). The increase in the number of "n–1" versus "1–1" orthologous lncRNAs for chicken, pig, and dog species is probably due to less accurate modeling of gene structures in these species compared to that in humans where transcript boundaries are validated by manual curation and 5′/3′ experimental supports (*e.g.* CAGE and polyA signals); the n gene would correspond to only one gene or some of them would actually 5′ or 3′ UTRs of neighboring protein-coding genes (Muret et al. 2019). Interestingly, the sum of the "1 to 1" and "n to 1" orthologous lncRNAs between each domesticated and human species is around 10% of the total lncRNAs in each species (Fig. 4B, right column) as reported in individuals studies of diverse species (Le Béguec et al. 2018; Kevin Muret et al. 2019; Breschi et al. 2017).

In conclusion, compared with the direct annotation of lncRNA gene structures, comparative genomic approaches allow strengthening the annotation of lncRNAs by providing insights into potentially functional lncRNAs related to a shared trait/disease, even though phylogenetic divergences should be considered for measuring the conservation of lncRNAs.

Long non-coding RNAs and transposons: towards long-read sequencing?

One of the most intriguing aspects of lncRNA biology lies in the observation that their sequences are highly enriched in transposable elements (TEs), that is, repetitive mobile elements capable of copying and moving into genomes. Briefly, TEs can be classified into two classes based on the mechanism by which they integrate into genomes. The first class, defined as retrotransposable elements, make use of a "copy-and-paste" strategy via the production of an intermediate RNA molecule, which is reverse transcribed into cDNA in order to be inserted into the genome. Usually, class 1 is subdivided into long terminal repeat (LTR) and non-LTR according to the biochemical mechanism of chromosomal integration, with non-LTR regrouping short and long interspersed nuclear elements (SINEs and LINEs). The second class of TEs, corresponding to DNA transposons, are

mobilised into genomes through a "cut-and-paste" strategy whereby a DNA intermediate is produced. In humans, more than 80% of lncRNAs overlap at least one annotated TE, with 40% of lncRNA sequences being derived from TEs (Kelley and Rinn 2012; Kapusta and Feschotte 2014). This led some authors to hypothesise that TEs are the functional domains of lncRNAs (Johnson and Guigo 2014). Indeed, it has recently been shown that specific repeat families can drive nuclear retention of lncRNAs in humans (Lubelsky and Ulitsky 2018; Carlevaro-Fita et al. 2019) or regulate mRNA translation (Zucchelli et al. 2015).

Regarding the 5 domesticated species studied in this review, the proportion of each reference assembly covered by TEs annotated by the RepeatMasker (http://www.repeatmasker.org) varies from 9.5% for chicken (galGal6) to 46.8% for the cow (bosTau9) (Fig. 5A). The lower proportion of TEs in the chicken genome could possibly be explained by the low copy numbers of SINE elements (< 10,000) compared with other mammals, such as humans (> 1,500,000) (Kapusta and Suh 2017). More specifically, SINE retrotransposons cover less than 0.1% of the chicken genome (7.6 Mb) as compared, for instance, to 10.5% (253 Mb) and 14.4% (359 Mb) for dog and pig genomes, respectively. When intersecting the annotations of lncRNAs and mobile genetic elements, between 23% of lncRNA transcripts for chicken and 84% for pigs are overlapped by at least one TE (Fig. 5B). In addition, when increasing the fraction of lncRNA transcript sequences that are overlapped by TEs, pig lncRNAs are still remarkably different from those of other mammals, with 41.1% and 18.7% of pig lncRNA sequences being composed of at least 5% and 10% of transposable elements, respectively (Fig. 5B). The inclusion of long-read transcriptomic data in the Ensembl-based annotation of pig lncRNAs has probably allowed a better reconstruction of lncRNA transcripts embedding repetitive elements such as TEs (See Fig. 5).

In line with this observation, recent transcriptome sequencing studies using long-read RNAseq (LR-RNAseq) promise to revolutionise annotation methods. Indeed, all reads from short-read RNAseq (SR-RNAseq) that are shorter than a specific repeat length will, by definition, not be uniquely assigned to one genome position, and thus would be considered as "multimapped". This can have a major impact on transcriptome reconstruction, especially for repeat-associated transcripts, such as lncRNAs. Steijger et al. showed that the best-performing method for reconstructing transcript models based on SR-RNAseq identified at most 21% of spliced transcripts in humans (Steijger et al. 2013). More recent studies involving the capture of lncRNAs followed by LR-RNAseq highlighted novel features for human and mouse lncRNA gene structures with (i) extensions of their 5′ and 3′ ends, (ii) similar splice length and exon count as in mRNAs (Lagarde et al. 2017), and (iii) near universal splicing of non-coding exons (Deveson et al. 2018). In addition to transcript structure, LR-RNAseq can allow the improved quantification of repeat-associated transcripts compared with SR-RNAseq (Sessegolo et al. 2019; Workman et al. 2019). Given that LR-RNA sequencing technologies represent an unfragmented vision of the transcriptome, they will more likely also facilitate gene reconstruction in domesticated species by direct exon/exon connectivity and read spanning repeats.

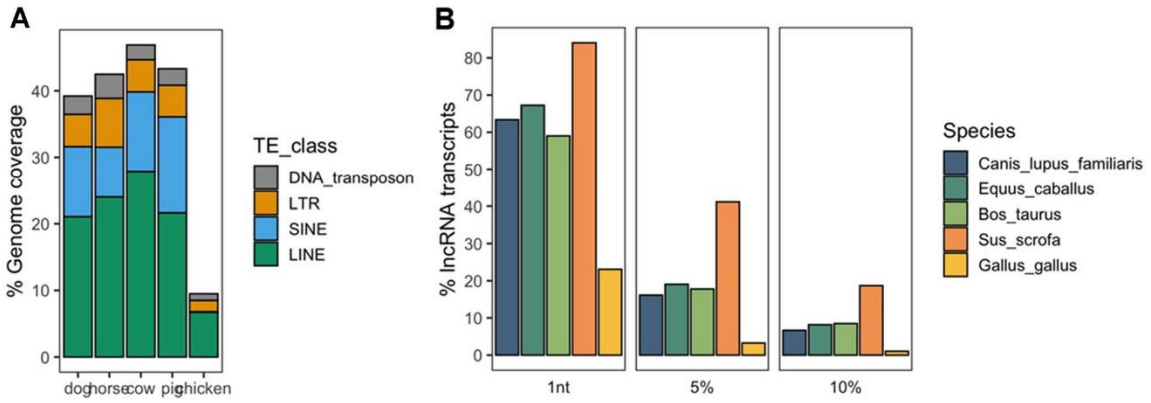


Fig. 5 Association between transposable elements (TEs) annotated by RepeatMasker and long non-coding RNAs annotated by Ensembl (v103) in 5 genome assemblies (canFam3, equCab3, bosTau9, susScr11 and galGal6). **A** Proportion of the genome covered by four TEs classes: LINEs, SINEs, LTRs, and DNA_transposons in green, blue, orange, and grey, respectively. **B** Proportion of Ensembl-based lncRNA transcripts overlapped by TEs for three fractions overlap (≥ 1 nucleotide, ≥ 5%, and ≥ 10% of the lncRNA sequence) in five domesticated species (dog, horse, cow, pig, and chicken, respectively, in dark green, orange, purple, pink, and light green)

Functions of long non-coding RNAs in domesticated species

As we have seen, lncRNA annotations have been associated with contrasted conditions, genotypes or GWAS hits (Table 2). However, as for human or model species (Bassett et al. 2014), assigning a functional mechanism to a lncRNA also remains a difficult task in domesticated species. Regarding GWAS, the first issue lies in identifying the causative variant in the GWAS region: in general, several polymorphisms being in linkage disequilibrium in the GWAS interval without the possibility to target the causative one(s) because of the low number of contrasted phenotypes which are observed (*i.e.* meiosis). When the mutation is located outside of a gene body, the second obstacle is to determine which gene is regulated by this polymorphism since the regulatory elements (*e.g.* enhancers) can act distantly from the targeted gene. Finally, the last difficulty is to validate the impact of the lncRNA gene (containing the polymorphism or regulated by this one) on the phenotype of interest. This last difficulty can be generalized to different observational levels such as animal, tissue, cell phenotypes. Thus, although tens of thousands of lncRNAs have been identified in the genomes of animals, their functions remain mostly unknown, irrespective of species. A review in 2019 reported that only 60 lncRNAs were involved in lipid metabolism despite the high number of lncRNAs identified in related tissues (e.g. liver or adipose tissue); these lncRNAs were mainly described in human or mouse, with only a precise described mode of action for a few of them (Muret et al. 2019). The main reasons probably stem from (i) an incomplete characterization of lncRNA isoforms and promoter sequences, (ii) a poor knowledge of the functionally important patches of lncRNA sequences, (iii) a lower expression level, and finally iv) multiple modes of action with cis or trans effect (Bassett et al. 2014).

In summary and as illustrated in the next paragraphs, only the function of a little number of lncRNAs has been elucidated in domesticated species.

GWAS hits involving lncRNAs

So far, there are only a few studies that have pinpointed lncRNAs located in GWAS intervals associated with a particular disease or trait. These studies have combined different approaches based on either genetic interval refinement using additional animals (and therefore meiosis) and/or molecular experiments to more deeply conclude the causative status of the lncRNA, although never with a formal demonstration with an in vivo experiment.

Concerning the dog species, Plassais et al. identified a ~ 1.5 Mb locus after GWAS involving hunting dog breeds affected by Human Sensory Autonomic Neuropathy (HSAN). After targeted DNA sequencing of the locus in four breeds, one exonic point mutation in an intergenic lncRNA termed *GDNF-AS* (transcribed in antisense orientation of the *GDNF* gene) was identified in affected dogs and absent in a panel of > 800 healthy dogs. By qRT-PCR analysis, a significant decrease of both the lincRNA and the mRNA expression levels was observed in specific tissues (*e.g.* dorsal root ganglia). In addition, gel shift assay (EMSA) revealed that the mutation significantly altered the binding of a transcription factor, altogether suggesting that *GNDF-AS* functions as an enhancer RNA (eRNA).

Concerning the livestock species, we can cite the callipyge (CLPG) locus responsible for muscle hypertrophy in sheep in which the *CLPG* mutation has been deeply studied and shown as interacting in trans between a maternally expressed repressor lncRNA, *MEG3* (alias *GTL2),* and its paternally expressed hypertrophy-promoting target, *DLK1* (Georges et al. 2003).

Another example concerns the Celtic Polled locus in cattle. Initially, a rather limited candidate region of 400 kb was identified by GWAS but contained numerous candidate polymorphisms. The study of new cases with versus without phenotype combined with different genotyping strategies allowed to reduce the number of candidate polymorphisms to a single one, the causal mutation (PC/c). The qRT-PCR analysis of the 7 genes located in the 500 kb upstream and downstream of the PC/c mutation revealed only one gene a differentially expressed between PC/p polled versus WT animals, a lincRNA without known function (Allais-Bonnet et al. 2013).

Other lncRNAs have been associated with a trait of interest by GWAS but these association studies require further investigations to confirm their phenotypic causality status because of the many SNPs in linkage disequilibrium. For instance, we can mention the lncRNAs *pouBW1* (Mei et al. 2016) or *pouMU1* (Ren et al. 2017a, b) related to chicken growth or the *lncRNA8138.1* related to reproductive traits.

Functional analysis by molecular biology approaches

Functional analysis by knock-out and knock-down

Validation of a single long non-coding RNA candidate Pioneer researchers studying specific lncRNAs have recycled methods initially developed for other classes of RNAs, such as tRNAs and mRNAs. To assign functions to lncRNAs, geneticists have successfully generated knock-outs (KO) or knock-downs (KD) of lncRNAs in cells or animal models (Knott and Doudna 2018). However, these target-

ing approaches have given rise to two main considerations regarding lncRNA specificities. Generating a lncRNA KO by deleting any exon without knowing its functional status could be risky. A more radical approach would be to delete the whole lncRNA gene or target the lncRNA promoters. In the last case, it is important to (i) verify that this promoter is not shared with another gene as in the case of bidirectional lncRNAs (Zhu et al. 2016), (ii) to evaluate the expression levels of neighbouring genes, and (iii) to perform rescue experiments.

A lncRNA depletion could be achieved using sequence-specific antisense oligonucleotides (ASO) able to target nuclear lncRNAs in contrast to small interfering RNAs (siRNA), thus efficiently knocking them down through the promotion of their RNAse H degradation (gapmers) (Crooke et al. 2021). The main pitfall relies on the efficient targeting of the lncRNA isoform of interest by short ASO (16–24 nucleotides) and could require preliminary experiments to determine the different transcript isoforms of the studied model.

Screening approaches To more systematically identify the functional role of lncRNAs, a screening approach might be sometimes attempted in parallel to high-throughput RNA-sequencing. CRISPR libraries for all human protein-coding genes (~20,000 genes) are available from non-profit companies (*e.g.* Addgene) for the performance of loss of function (CRISPR KO), gain-of-function (CRISPR activator, CRISPRa), or mRNA knockdown studies via CRISPR inhibition (CRISPRi) at a modest cost (< 500 €). These libraries, containing 3–10 single guide RNAs (sgRNAs) per targeted transcript, have been validated in various studies (Konermann et al. 2015; Joung et al. 2017). However, CRISPR KO libraries seem inappropriate for lncRNAs, as the functional domain(s) of lncRNAs have not been yet clearly identified. In contrast, CRISPRa and CRISPRi strategies (Liu et al. 2017a, b; Esposito et al. 2019) could efficiently modulate the expression (up- or downregulate) of lncRNAs; however, 2 main limitations need to be mentioned. First, the single guide RNA (sgRNA) libraries have been designed from lncRNA databases, such as Ensembl or GENCODE, built on models reconstructed from RNAseq data of different cell types or differentiation states and therefore not specific for a given cell type/tissue; thus, many sgRNA might not be functional in the studied cell model given the high tissue- and condition-specific feature of lncRNAs. Second, the design of a sgRNA library might be sometimes hazardous because of the imperfect knowledge of lncRNA promoter regions, despite the recent advancements in 5′ end annotation in human, dog, and chicken (Hon et al. 2017). To the best of our knowledge, such CRISPR libraries are not yet available for domesticated species.

Even if these two strategies (KO & KD) are correctly evaluated, other complementary experiments would still be required to establish the mode of action of these lncRNAs.

Long non-coding RNA interacting partners

The functions of lncRNAs have been previously reviewed (Quinn and Chang 2016; Gil and Ulitsky 2019; Statello et al. 2021). Their functional mechanisms are diverse, including lncRNAs that act as scaffolds, decoys, or signals. In addition, they can act by regulating in both cis or trans (Ulitsky and Bartel 2013; Geisler and Coller 2013).

Interacting partner detection Numerous methods have been developed to identify the interactions of lncRNAs with either RNA, DNA, or proteins (Goff and Rinn 2015). Despite their differences, the principle is often the same requiring, the enrichment of lncRNA partners using lncRNA precipitation. Most groups performed lncRNA precipitation using short oligonucleotides coupled to biotin. Based on complementary base-pairing, ribonucleotide complex-associated to the biotinylated ASO were purified via streptavidin beads followed by stringent washes. The identity of the partner was revealed using sequencing analyses (RNA or DNA) or spectrometry (proteins). As with all enrichment experiments, false positives and false negatives are inherent to these approaches, rendering the performance of validation experiments a crucial step. When an lncRNA-interactant is identified, complementary experiments are needed to validate the domain of lncRNA interacting with a protein or an RNA or a DNA sequence. Depending on the lncRNA-interactant nature, different experiments can be envisaged.

Interaction domain identification While robust, the conventional protein immunoprecipitation followed by lncRNA detection (RT-qPCR) requires an efficient crosslinking between the lncRNA and the protein (before IP), which is not always possible in animal models. A biotinylated short-RNA complementary to the RNA interactant is usually used as a bait for the successful purification and detection of lncRNA-RNA interactions using streptavidin beads. Similar approaches are used for DNA, but involve an efficient DNA fragmentation or partial digestion using recombinant restriction enzymes (Chu et al. 2015).

Validation of the interacting domain by inhibiting interaction An elegant detection strategy works by preventing the binding between the candidate partner and the studied lncRNA. This can be achieved by protecting or deleting the interacting domain of the lncRNA. The second strategy is based on the prime-editing approach published in 2019 (Anzalone et al. 2019). This CRISPR 3.0 method allows researchers to rewrite the DNA sequence encoding the

Table 3 LncRNA studies associated with in vitro functional analyses for livestock species

lncRNA name	lncRNA impact	Cellular model	Strategy	Year (Refs)
A. Chicken				
MHM	Embryonic development Sex determination	Egg (0-day blastoderms)	OverEx	2012 (Roeszler et al. 2012)
B. Cow				
ADNCR	Impact on SIRT1 by competing with miR-204 as a ceRNA to regulate adipogenesis	HEK293T,HEK293A & ADSC cells	OverEx KD by siRNA	2016 (Li et al. 2016)
LncRNA candidate 1	Embryonic developmental rates	Cattle matured oocytes	KD by siRNA	2015 (Caballero et al. 2014)
H19	Differentiation of satellite cells. Blocking of the Sirt1/FoxO1 pathway during myogenesis	C_2C_{12} cells & satellite cells (from adult cattle muscle)	OverEx KD by pLenti-NTC interference vector	2017 (Xu et al. 2017)
lnc403	Inhibit myogenic differentiation of bovine skeletal muscle satellite cells Negatively regulated gene Myf6 and positively regulated protein KRAS	Satellite cells (from foetal bovine muscle)	OverEx KD by siRNA	2020 (Zhang et al. 2020)
IGF2 AS	Promote proliferation and differentiation of bovine myoblasts through various pathways	Myoblasts (from foetal bovine muscle)	OverEx KD by siRNA	2020 (Song et al. 2020)
C. Pig				
lncIMF4	Associated with adipogenesis and effect in intramuscular preadipocyte proliferation and differentiation	Intramuscular preadipocytes (from 2 pig breeds)	KD by siRNA	2020 (Sun et al. 2020)
TCONS_00815878	Decreasing of Myod, MyoG and MyHC such as glycolysis and pyruvate metabolism which are related to skeletal muscle satellite cell differentiation	Skeletal muscle satellite cells	KD by ASO	2019 (Huang et al. 2019)
XLOC-2222497	Regulate AKR1C1 and progesterone metabolism	Endometrial cells	OverEx KD by ASO	2020 (Su et al. 2020)

KD knock-down, *OverEx* overexpression

lncRNA or the putative partner. To date, this method is probably the most appropriate for studying lncRNA domains and functions because the experiments are based on the normal expression level of the lncRNA. More specifically, experiments do not require the overexpression of the lncRNA or its putative partner. Although this approach is clever, designing an efficient prime-editing sgRNA (pegRNA) is difficult (Lin et al. 2020a, b; Marzec and Hensel 2020). Given that the efficiency of a pegRNA varies between 0.1% and 50%, many clones must be sequenced before the identification of the correct edited clone (*i.e.* homozygous edition).

Examples in domesticated animals

As described above, RNA interaction experiments as knock-out and knock-down using CRISPR tools coupled to ASOs are well suited to elucidate the functions of lncRNAs both in vitro and in vivo. Concerning the in vitro studies (*i.e.* using a cellular system), while overexpression and knock-down experiments are reported in domesticated species for protein-coding genes, this type of studies is less frequent for lncRNAs. Table 3 provides a few studies associated with in vitro functional analyses of lncRNA for livestock species. We can note that some studies start to use ASO sequences which are more efficient to deplete the target lncRNA than siRNA. Concerning the in vivo studies allowing to formally validate the impact of a gene mutation on a phenotype, they are still limited for protein-coding genes. We can cite the disruption of the *CD163* gene in pigs by CRISPR conferring resistance to PRRSV infection, the activation of the *MSTN* gene (myostatin) in sheep and cow resulting in meat production improvement (for review, see (Menchaca et al. 2020) or the correction of muscular dystrophies in dogs using CRISPR targeting the *DMD* gene (dystrophin) (Amoasii

et al. 2018). To the best of our knowledge, such studies do not yet exist for lncRNAs.

Conclusion/perspectives

Domestic animals have been selectively bred by humans during thousands of years for cultural or economic reasons. Consequently, they provide an almost infinite space of desired phenotypes involving genomic variations in protein-coding and non-coding elements. Although the former has been studied for a long time, the importance of long non-coding RNAs has only been investigated recently in human and model organisms, and even more recently in domesticated animals. Despite the democratization of short-read RNAseq combined with efficient bioinformatic programs to manage these data, we showed that lncRNA annotations in domesticated animals are far from complete as compared to human or mouse, both in terms of number of gene loci and alternative isoforms. Moreover, the catalogues of lncRNAs available in public resources display a very low overlap. As we have seen, this can mainly be explained by the specific features of lncRNAs (high tissue-specificity, low expression levels, high repeat content, …) and the limited number of RNAseq samples used for generating these catalogues, even for dedicated annotation resources such as Ensembl or NCBI/RefSeq. Furthermore, the diverse computational solutions used by these resources probably impact the number of shared lncRNAs, by defining dissimilar gene boundaries (at the transcriptome reconstruction step) or by misclassifying transcript biotypes (at the coding-potential assessment step).

In order to leverage the importance of lncRNAs in animal models and evaluate their functionality, several complementary directions could be envisaged to increase the completeness of the annotations and to provide more accurate catalogues of lncRNAs. The first one relies on exploiting and combining the wealth of public RNASeq data available in public repositories (SRA/ENA) in order to include as many as possible tissues, physiological/pathological stages and environmental conditions. Although feasible in theory, this requires efficient programs and large computational infrastructures to regularly cope with the thousands of data now available for domesticated species and to carefully version each newly produced catalogues (Seal et al. 2020).

As mentioned previously (Steijger et al. 2013), one of the major bottlenecks in the bioinformatic process of annotating gene models can be related to the transcript reconstruction step *i.e.* the process of connecting multiple exons into correct spliced isoforms. The growing interest in long-read RNA sequencing, provided by technologies such as ONT or PacBio, will likely facilitate the reconstruction of full-length non-coding (and coding) gene models for domesticated species in the near future. Yet, these technologies still produce shallow sequencing depths compared to short-read RNAseq. This could be an issue for lowly expressed transcripts such as lncRNAs although capture strategies followed by LR-RNAseq have been recently applied with success in human and mouse (Lagarde et al. 2017).

The availability of these catalogues of lncRNAs in domesticated species, even if not perfect, has allowed researchers to include these new types of regulatory genes in their studies, by showing some of these lncRNAs to be differentially expressed across treatments, conditions, or genotypes. To go further on some lncRNAs of interest, it is important to keep in mind that multiple evidence should be considered to assess lncRNA functionality in domesticated animals. The identification of an orthologous lncRNA, by sequence or positional conservation, in human databases is a good proxy for its real existence but would involve that the phenotype of interest is evolutionary conserved between the studied domesticated species and human. While information has been gained about the evolution of lncRNAs across distantly related species through large-scale comparative transcriptomic studies, very little is known regarding the conservation of lncRNAs at smaller time-scale (e.g. between populations within a species). The genetic architecture of domesticated species, with homogeneous breed/population structure and potential large-scale phenotypic data, represent ideal models for dissecting the impact of the non-coding genome on a breed-associated trait. The combination of exhaustive/accurate lncRNA genomic maps with standardized functional technologies (e.g. ASO or CRISPR) represent a prerequisite to assess lncRNA functionality and will pave the way to decipher the role of these enigmatic transcripts in the phenotypes of domesticated animals.

Acknowledgements Authors would like to thank people from the Genetics & Genomics Team, the Canine Genetics Team (Benoit Hédan), Jocelyn Plassais for helpful comments and the Genouest bioinformatic core facility (https://www.genouest.org).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Adetula AA, Lantao Gu, Nwafor CC, Xiaoyong Du, Zhao S, Li S (2018) Transcriptome sequencing reveals key potential long non-coding RNAs related to duration of fertility trait in the utero-ovaginal junction of egg-laying hens. Sci Rep 8(1):13185. https://doi.org/10.1038/s41598-018-31301-z

Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Banet JF et al (2016) The ensembl gene annotation system. Database. https://doi.org/10.1093/database/baw093

Alexandre PA, Reverter A, Berezin RB, Porto-Neto LR, Ribeiro G, Santana MHA, Ferraz JBS, Fukumasu H (2020) Exploring the regulatory potential of long non-coding rna in feed efficiency of indicine cattle. Genes. https://doi.org/10.3390/genes11090997

Allais-Bonnet A, Grohs C, Medugorac I, Krebs S, Djari A, Graf A, Fritz S et al (2013) Novel insights into the bovine polled phenotype and horn ontogenesis in bovidae. PLoS ONE 8(5):e63512. https://doi.org/10.1371/journal.pone.0063512

Amoasii L, Hildyard JCW, Li H, Sanchez-Ortiz E, Mireault A, Caballero D, Harron R et al (2018) Gene editing restores dystrophin expression in a canine model of duchenne muscular dystrophy. Science 362(6410):86–91. https://doi.org/10.1126/science.aau1549

Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E et al (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. Genome Biol 16(1):57. https://doi.org/10.1186/s13059-015-0622-4

Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ et al (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 576(7785):149–157. https://doi.org/10.1038/s41586-019-1711-4

Appelbaum T, Murgiano L, Becker D, Santana E, Aguirre GD (2020) Candidate genetic modifiers for RPGR retinal degeneration. Invest Ophthalmol Vis Sci 61(14):20. https://doi.org/10.1167/iovs.61.14.20

Bakhtiarizadeh MR, Hosseinpour B, Arefnezhad B, Shamabadi N, Salami SA (2016) In silico prediction of long intergenic non-coding RNAs in sheep. Genome 59(4):263–275. https://doi.org/10.1139/gen-2015-0141

Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, Ephrussi A et al (2014) Considerations when investigating LncRNA function in vivo. Elife 3:e03058. https://doi.org/10.7554/eLife.03058

Bazzini AA, Johnstone TG, Christiano R, MacKowiak SD, Obermayer B, Fleming ES, Vejnar CE et al (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J 33(9):981–993. https://doi.org/10.1002/embj.201488411

Béguec Le, Céline VW, Lagoutte L, Cadieu E, Botherel N, Hédan B, De Brito C et al (2018) Characterisation and functional predictions of canine long non-coding RNAs. Sci Rep 8(1):13444. https://doi.org/10.1038/s41598-018-31770-2

Beiki H, Liu H, Huang J, Manchanda N, Nonneman D, Smith TPL, Reecy JM, Tuggle CK (2019) Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-Seq data. BMC Genomics 20(1):344. https://doi.org/10.1186/s12864-019-5709-y

Breschi A, Gingeras TR, Guigó R (2017) Comparative transcriptomics in human and mouse. Nat Rev Genet 18(7):425–440. https://doi.org/10.1038/nrg.2017.19

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW et al (2014) Diversity and dynamics of the drosophila transcriptome. Nature 512(7515):393–399. https://doi.org/10.1038/nature12962

Bu D, Luo H, Jiao F, Fang S, Tan C, Liu Z, Zhao Y (2015) Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. Sci China Life Sci 58(8):787–798. https://doi.org/10.1007/s11427-015-4881-9

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A et al (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47(D1):D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bush SJ, Muriuki C, McCulloch MEB, Farquhar IL, Clark EL, Hume DA (2018) Cross-species inference of long non-coding RNAs greatly expands the ruminant transcriptome. Genet Sel Evol 50(1):20. https://doi.org/10.1186/s12711-018-0391-0

Caballero J, Gilbert I, Fournier E, Gagné D, Scantland S, Macaulay A, Robert C (2014) Exploring the function of long non-coding RNA in the development of bovine early embryos. Reprod Fertil Dev 27(1):40–52. https://doi.org/10.1071/RD14338

Cabili M, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25(18):1915–1927. https://doi.org/10.1101/gad.17446611

Cai B, Li Z, Ma M, Wang Z, Han P, Abdalla BA, Nie Q, Zhang X (2017) LncRNA-Six1 encodes a micropeptide to activate Six1 in cis and is involved in cell proliferation and muscle growth. Front Physiol 8:230. https://doi.org/10.3389/fphys.2017.00230

Carlevaro-Fita J, Polidori T, Das M, Navarro C, Zoller TI, Johnson R (2019) Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. Genome Res 29(2):208–222. https://doi.org/10.1101/gr.229922.117

Carninci P (2005) The transcriptional landscape of the mammalian genome. Science 309(5740):1559–1563. https://doi.org/10.1126/science.1112014

Cascione L, Giudice L, Ferraresso S, Marconato L, Giannuzzi D, Napoli S, Bertoni F, Giugno R, Aresu L (2019) Long non-coding RNAs as molecular signatures for canine B-cell lymphoma characterization. Non-Coding RNA. https://doi.org/10.3390/ncrna5030047

Che T, Li D, Jin L, Yuhua Fu, Liu Y, Liu P, Wang Y et al (2018) Long non-coding RNAs and MRNAs profiling during spleen development in pig. PLoS ONE 13(3):e0193552. https://doi.org/10.1371/journal.pone.0193552

Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, Hanna JH, Regev A, Garber M (2016) Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol 17(1):19. https://doi.org/10.1186/s13059-016-0880-9

Chen L, Zhang T, Zhang S, Huang J, Zhang G, Xie K, Wang J, Haiqing W, Dai G (2019) Identification of long non-coding RNA-associated competing endogenous RNA network in the differentiation of chicken preadipocytes. Genes. https://doi.org/10.3390/genes10100795

Choi J-Y, Shin D, Lee H-J, Jae-Don Oh (2019) Comparison of long noncoding RNA between muscles and adipose tissues in hanwoo beef cattle. Anim Cells Syst 23(1):50–58. https://doi.org/10.1080/19768354.2018.1512522

Chu Ci, Spitale RC, Chang HY (2015) Technologies to probe functions and mechanisms of long noncoding RNAs. Nat Struct Mol Biol 22(1):29–35. https://doi.org/10.1038/nsmb.2921

Crooke ST, Baker BF, Crooke RM, Liang X-H (2021) Antisense technology: an overview and prospectus. Nat Rev Drug Dis. https://doi.org/10.1038/s41573-021-00162-z

Dahlgren AR, Scott EY, Mansour T, Hales EN, Ross P, Kalbfleisch TS, MacLeod JN, Petersen JL, Bellone RR, Finno CJ (2020) Comparison of poly-A+ selection and RRNA depletion in detection of LncRNA in two equine tissues using RNA-Seq. Non-Coding RNA. https://doi.org/10.3390/ncrna6030032

Dai M, Feng M, Xie T, Zhang X (2019) Long non-coding RNA and MicroRNA profiling provides comprehensive insight into non-coding RNA involved host immune responses in ALV-J-infected chicken primary macrophage. Dev Comp Immunol 100(November):103414. https://doi.org/10.1016/j.dci.2019.103414

Darbellay Fabrice, Necsulea Anamaria (2020) "Comparative transcriptomics analyses across species, organs, and developmental stages reveal functionally constrained LncRNAs." Edited by Amanda Larracuente. Mol Biol Evol 37(1):240–59. https://doi.org/10.1093/molbev/msz212

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G et al (2012) The GENCODE v7 catalogue of human long non-coding RNAs: analysis of their structure, evolution and expression. Genome Res 22:1775–1789. https://doi.org/10.1101/gr.132159.111

Deveson IW, Brunck ME, Blackburn J, Nielsen LK, Mattick JS, Mercer TR, Tseng E et al (2018) Universal alternative splicing of noncoding exons. Cell Syst 6(2):245-255.e5. https://doi.org/10.1016/j.cels.2017.12.005

Ding H, Zhao H, Zhao X, Qi Y, Wang X, Huang D (2021) Analysis of histology and long noncoding RNAs involved in the rabbit hair follicle density using RNA sequencing. BMC Genomics 22(1):89. https://doi.org/10.1186/s12864-021-07398-4

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A et al (2012) Landscape of transcription in human cells. Nature 489(7414):101–108. https://doi.org/10.1038/nature11233

Djebali S, Wucher V, Foissac S, Hitte C, Corre EE, Derrien T (2017) Bioinformatics pipeline for transcriptome sequencing analysis. Methods in Molecular Biology (Clifton, N.J.) 1468:201–19. https://doi.org/10.1007/978-1-4939-4035-6_14

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635

Duret L (2006) The xist RNA gene evolved in Eutherians by pseudogenization of a protein-coding gene. Science 312(5780):1653–1655. https://doi.org/10.1126/science.1126316

Esposito R, Bosch N, Lanzós A, Polidori T, Pulido-Quetglas C, Johnson R (2019) Hacking the cancer genome: profiling therapeutically actionable long non-coding RNAs using CRISPR-Cas9 screening. Cancer Cell 35(4):545–557. https://doi.org/10.1016/j.ccell.2019.01.019

Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Pérez-Enciso M (2011) Exploring the gonad transcriptome of two extreme male pigs with RNA-Seq. BMC Genomics 12(November):552. https://doi.org/10.1186/1471-2164-12-552

Fang M, Yang Yi, Wang N, Wang A, He Y, Wang J, Jiang Y, Deng Z (2019) Genome-wide analysis of long non-coding RNA expression profile in porcine circovirus 2-infected intestinal porcine epithelial cell line by RNA sequencing. PeerJ 7:e6577. https://doi.org/10.7717/peerj.6577

Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, Esquerré D et al (2019) Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biol 17(1):108. https://doi.org/10.1186/s12915-019-0726-5

Food and Agriculture Organization of the United Nations. 2021. *FAOSTAT, Livestock primary*. FAO. http://www.fao.org/faostat/en/#data/QL. Accessed 14 Jul 2021

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM et al (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47:767. https://doi.org/10.1093/nar/gky955

Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A et al (2016) Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science 352(6290):1228–1231. https://doi.org/10.1126/science.aaf3161

Gao Y, Li S, Lai Z, Zhou Z, Fei Wu, Huang Y, Lan X, Lei C, Chen H, Dang R (2019) Analysis of long non-coding RNA and MRNA expression profiling in immature and mature bovine (Bos Taurus) testes. Front Genet 10:646. https://doi.org/10.3389/fgene.2019.00646

Geisler S, Coller J (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol 14(11):699–712. https://doi.org/10.1038/nrm3679

Georges M, Charlier C, Cockett N (2003) The callipyge locus: evidence for the trans interaction of reciprocally imprinted genes. Trends in Genetics 19(5):248–252. https://doi.org/10.1016/S0168-9525(03)00082-9

Gil N, Ulitsky I (2019) Regulation of gene expression by cis-acting long non-coding RNAs. Nat Rev Genetics. https://doi.org/10.1038/s41576-019-0184-5

Goff LA, Rinn JL (2015) Linking RNA biology to LncRNAs. Genome Res 25(10):1456–1465. https://doi.org/10.1101/gr.191122.115

Guo J-C, Fang S-S, Yang Wu, Zhang J-H, Chen Y, Liu J, Bo Wu et al (2019) CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. Nucleic Acids Res 47(W1):W516–W522. https://doi.org/10.1093/nar/gkz400

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G et al (2011) LincRNAs Act in the circuitry controlling pluripotency and differentiation. Nature 477(7364):295–300. https://doi.org/10.1038/nature10398

He Y, Ding Yi, Zhan F, Zhang H, Han Bo, Gangqing Hu, Zhao K et al (2015) The conservation and signatures of LincRNAs in marek's disease of chicken. Sci Rep 5(October):15184. https://doi.org/10.1038/srep15184

Hédan B, Cadieu É, Rimbault M, Vaysse A, Dufaure C, de Citres P, Devauchelle NB et al (2021) Identification of common predisposing loci to hematopoietic cancers in four dog breeds. PLoS Genet 17(4):e1009395. https://doi.org/10.1371/journal.pgen.1009395

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep 11(7):1110–1122. https://doi.org/10.1016/j.celrep.2015.04.023

Hezroni H, Perry R-T, Meir Z, Housman G, Lubelsky Y, Ulitsky I (2017) A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. Genome Biol 18(1):162. https://doi.org/10.1186/s13059-017-1293-0

Hill WG (2016) Is continued genetic improvement of livestock sustainable? Genetics 202(3):877–881. https://doi.org/10.1534/genetics.115.186650

Hitte C, Le Béguec C, Cadieu E, Wucher V, Primot A, Prouteau A, Botherel N et al (2019) Genome-wide analysis of long non-coding RNA profiles in canine oral melanomas. Genes 10(6):477. https://doi.org/10.3390/genes10060477

Hoeppner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, Sundström G et al (2014) An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS ONE. https://doi.org/10.1371/journal.pone.0091172

Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJLL, Gough J, Denisenko E et al (2017) An atlas of human long non-coding RNAs with accurate 5′ ends. Nature 543(7644):199–204. https://doi.org/10.1038/nature21374

Hong L, Qun Hu, Zang X, Xie Y, Zhou C, Zou X, Li Y et al (2020) Analysis and screening of reproductive long non-coding RNAs through genome-wide analyses of goat endometrium during the

Li D, Li F, Jiang K, Zhang M, Han R, Jiang R, Li Z et al (2019) Integrative analysis of long noncoding RNA and MRNA reveals candidate LncRNAs responsible for meat quality at different physiological stages in gushi chicken. PLoS ONE 14(4):e0215006. https://doi.org/10.1371/journal.pone.0215006

Li Q, Qiao J, Zhang Z, Shang X, Chu Z, Yajuan Fu, Chu M (2020) Identification and analysis of differentially expressed long noncoding RNAs of Chinese holstein cattle responses to heat stress. Anim Biotechnol 31(1):9–16. https://doi.org/10.1080/10495398.2018.1521337

Li R, Li B, Jiang A, Cao Y, Hou L, Zhang Z, Zhang X, Liu H, Kim K-H, Wangjun W (2020) Exploring the LncRNAs related to skeletal muscle fiber types and meat quality traits in pigs. Genes. https://doi.org/10.3390/genes11080883

Li H, Cui P, Xue Fu, Zhang L, Yan W, Zhai Y, Lei C, Wang H, Yang X (2021) Identification and analysis of long non-coding RNAs and MRNAs in chicken macrophages infected with avian infectious bronchitis coronavirus. BMC Genomics 22(1):67. https://doi.org/10.1186/s12864-020-07359-3

Lian Zhiquan, Zou Xian, Han Yinru, Deng Ming, Sun Baoli, Guo Yongqing, Zhou Lei, Liu Guangbin, Liu Dewu, Li Yaokun (2020) Role of MRNAs and long non-coding RNAs in regulating the litter size trait in Chuanzhong black goats. Reprod Domest Anim = Zuchthygiene 55(4):486–95. https://doi.org/10.1111/rda.13642

Liang G, Yang Y, Li H, Yu H, Li X, Tang Z, Li K (2018) LncRNAnet: a comprehensive sus scrofa LncRNA database. Anim Genet 49(6):632–635. https://doi.org/10.1111/age.12720

Lin Q, Zong Y, Xue C, Wang S, Jin S, Zhu Z, Wang Y et al (2020) Prime genome editing in rice and wheat. Nat Biotechnol 38(5):582–585. https://doi.org/10.1038/s41587-020-0455-x

Lin Y, Yang J, He D, Li X, Li J, Tang Yi, Diao Y (2020) Differently expression analysis and function prediction of long non-coding RNAs in duck embryo fibroblast cells infected by duck tembusu virus. Front Immunol 11:1729. https://doi.org/10.3389/fimmu.2020.01729

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069):803–819. https://doi.org/10.1038/nature04338

Liu Y, Sun Y, Li Y, Bai H, Xue F, Songshan Xu, Hong Xu, Shi L, Yang N, Chen J (2017b) Analyses of long non-coding RNA and MRNA profiling using RNA sequencing in chicken testis with extreme sperm motility. Sci Rep 7(1):9055. https://doi.org/10.1038/s41598-017-08738-9

Liu L, Xiao Q, Gilbert ER, Cui Z, Zhao X, Wang Y, Yin H, Li D, Zhang H, Zhu Q (2018) Whole-transcriptome analysis of atrophic ovaries in broody chickens reveals regulatory pathways associated with proliferation and apoptosis. Sci Rep 8(1):7231. https://doi.org/10.1038/s41598-018-25103-6

Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ et al (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science 355(6320):7111. https://doi.org/10.1126/science.aah7111

Lu C, Xing Y, Cai H, Shi Y, Liu J, Huang Y (2019) Identification and analysis of long non-coding RNAs in response to H5N1 influenza viruses in duck (*Anas Platyrhynchos*). BMC Genomics 20(1):36. https://doi.org/10.1186/s12864-018-5422-2

Lubelsky Y, Ulitsky I (2018) Sequences enriched in alu repeats drive nuclear localization of long RNAs in human cells. Nat Publ Group. https://doi.org/10.1038/nature25757

Marzec M, Hensel G (2020) Prime editing: game changer for modifying plant genomes. Trends Plant Sci 25(8):722–724. https://doi.org/10.1016/j.tplants.2020.05.008

Mattick JS, Rinn JL (2015) Discovery and annotation of long noncoding RNAs. Nat Struct Mol Biol 22(1):5–7. https://doi.org/10.1038/nsmb.2942

Mei X, Kang X, Liu X, Jia L, Li H, Li Z, Jiang R (2016) Identification and SNP association analysis of a novel gene in chicken. Anim Genet 47(1):125–127. https://doi.org/10.1111/age.12387

Menchaca A, Dos Santos-Neto PC, Mulet AP, Crispo M (2020) CRISPR in livestock: from editing to printing. Theriogenology 150(July):247–254. https://doi.org/10.1016/j.theriogenology.2020.01.063

Miao Z, Wang S, Zhang J, Wei P, Guo L, Liu D, Wang Y, Shi M (2018) Identification and comparison of long non-conding RNA in Jinhua and landrace pigs. Biochem Biophys Res Commun 506(3):765–771. https://doi.org/10.1016/j.bbrc.2018.06.028

Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, Désert C et al (2017) Long noncoding RNA repertoire in chicken liver and adipose tissue. Genet Sel Evol 49(1):6. https://doi.org/10.1186/s12711-016-0275-0

Muret K, Désert C, Lagoutte L, Boutin M, Gondret F, Zerjal T, Lagarrigue S (2019) Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. BMC Genomics 20(1):882. https://doi.org/10.1186/s12864-019-6093-3

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H (2014) The evolution of LncRNA repertoires and expression patterns in tetrapods. Nature 505(7485):635–640. https://doi.org/10.1038/nature12943

Ning C, Ma T, Silu Hu, Zhongxian Xu, Zhang Pu, Zhao X, Wang Y et al (2020) Long non-coding RNA and MRNA profile of liver tissue during four developmental stages in the chicken. Front Genet 11:574. https://doi.org/10.3389/fgene.2020.00574

Nolte W, Weikard R, Brunner RM, Albrecht E, Hammon HM, Reverter A, Kühn C (2019) Biological network approach for the identification of regulatory long non-coding RNAs associated with metabolic efficiency in cattle. Front Genet 10:1130. https://doi.org/10.3389/fgene.2019.01130

Noviello TMR, Di Liddo A, Ventola GM, Spagnuolo A, D'Aniello S, Ceccarelli M, Cerulo L (2018) Detection of long non-coding RNA homology, a comparative study on alignment and alignment-free metrics. BMC Bioinform 19(1):407. https://doi.org/10.1186/s12859-018-2441-6

O'Leary NA, Wright MW, Rodney Brister J, Ciufo S, Haddad D, McVeigh R, Rajput B et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733-745. https://doi.org/10.1093/nar/gkv1189

Ostrander EA, Wayne RK, Freedman AH, Davis BW (2017) Demographic history, selection and functional diversity of the canine genome. Nat Rev Genet 18(12):705–720. https://doi.org/10.1038/nrg.2017.67

Ouyang Q, Shenqiang H, Wang G, Jiwei H, Zhang J, Li L, Bo H et al (2020) Comparative transcriptome analysis suggests key roles for 5-hydroxytryptamlne receptors in control of goose egg production. Genes. https://doi.org/10.3390/genes11040455

Özdemir S, Altun S (2020) Genome-wide analysis of MRNAs and LncRNAs in mycoplasma bovis infected and non-infected bovine mammary gland tissues. Mol Cell Probes 50(April):101512. https://doi.org/10.1016/j.mcp.2020.101512

Peng Y, Chang Li, Wang Y, Wang R, Lulu Hu, Zhao Z, Geng L et al (2019) Genome-wide differential expression of long noncoding RNAs and MRNAs in ovarian follicles of two different chicken breeds. Genomics 111(6):1395–1403. https://doi.org/10.1016/j.ygeno.2018.09.012

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015) Stringtie enables improved reconstruction of a

and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515. https://doi.org/10.1038/nbt.1621

Ulitsky I, Bartel DP (2013) XLincRNAs: genomics, evolution, and mechanisms. Cell 154(1):26–46. https://doi.org/10.1016/j.cell.2013.06.020

van Steenbeek FG, Hytönen MK, Leegwater PaJ, Lohi H (2016) The canine era: the rise of a biomedical model. Anim Genet 47(5):519–527. https://doi.org/10.1111/age.12460

Vancura A, Lanzós A, Bosch-Guiteras N, Esteban MT, Gutierrez AH, Haefliger S, Johnson R (2021) Cancer LncRNA census 2 (CLC2): an enhanced resource reveals clinical features of cancer LncRNAs. NAR Cancer. https://doi.org/10.1093/narcan/zcab013

Verma A, Jiang Y, Wei Du, Fairchild L, Melnick A, Elemento O (2015) Transcriptome sequencing reveals thousands of novel long non-coding RNAs in B cell lymphoma. Genome Med 7(November):110. https://doi.org/10.1186/s13073-015-0230-7

Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63. https://doi.org/10.1038/nrg2484

Wang Y, Xue S, Liu X, Liu H, Tao Hu, Qiu X, Zhang J, Lei M (2016) Analyses of long non-coding RNA and MRNA profiling using RNA sequencing during the pre-implantation phases in pig endometrium. Sci Rep 6(January):20238. https://doi.org/10.1038/srep20238

Wang J, Hua L, Chen J, Zhang J, Bai X, Gao B, Li C et al (2017) Identification and characterization of long non-coding RNAs in subcutaneous adipose tissue from castrated and intact full-sib pair huainan male pigs. BMC Genomics 18(1):542. https://doi.org/10.1186/s12864-017-3907-z

Wang G-Z, Kun Du, Shen-Qiang Hu, Chen S-Y, Jia X-B, Ming-Cheng Cai Yu, Shi JW, Lai S-J (2018) Genome-wide identification and characterization of long non-coding RNAs during postnatal development of rabbit adipose tissue. Lipids Health Dis 17(1):271. https://doi.org/10.1186/s12944-018-0915-1

Wang X, Yang C, Guo F, Zhang Y, Zhihua Ju, Jiang Q, Zhao X et al (2019a) Integrated analysis of MRNAs and long noncoding RNAs in the semen from holstein bulls with high and low sperm motility. Sci Rep 9(1):2092. https://doi.org/10.1038/s41598-018-38462-x

Wang Z, Yang Y, Li S, Li K, Tang Z (2019b) Analysis and comparison of long non-coding RNAs expressed in the ovaries of meishan and yorkshire pigs. Anim Genet 50(6):660–669. https://doi.org/10.1111/age.12849

Wang J-J, Niu M-H, Zhang T, Shen W, Cao H-G (2020) Genome-wide network of LncRNA-MRNA during ovine oocyte development from germinal vesicle to metaphase II in vitro. Front Physiol 11:1019. https://doi.org/10.3389/fphys.2020.01019

Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, Chow W et al (2020) An improved pig reference genome sequence to enable pig genetics and genomics research. GigaScience 9(6):1–14. https://doi.org/10.1093/gigascience/giaa051

Washietl S, Kellis M, Garber M (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res 24(4):616–628. https://doi.org/10.1101/gr.165035.113

Weikard R, Demasius W, Kuehn C (2017) Mining long noncoding RNA in livestock. Anim Genet 48(1):3–18. https://doi.org/10.1111/age.12493

Weikard R, Hadlich F, Hammon HM, Frieten D, Gerbert C, Koch C, Dusel G, Kuehn C (2018) Long noncoding RNAs are associated with metabolic and cellular processes in the jejunum mucosa of pre-weaning calves in response to different diets. Oncotarget 9(30):21052–69. https://doi.org/10.1832/oncotarget.24898

WibergHalligan RADL, Ness RW, Necsulea A, Kaessmann H, Keightley PD (2015) Assessing recent selection and functionality at long noncoding RNA loci in the mouse genome. Genome Biol Evol 7(8):2432–2444. https://doi.org/10.1093/gbe/evv155

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC et al (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. https://doi.org/10.1038/s41592-019-0617-2

Wu S, Liu Y, Guo W, Cheng Xi, Ren X, Chen Si, Li X, Duan Y, Sun Q, Yang X (2018) Identification and characterization of long noncoding RNAs and MRNAs expression profiles related to postnatal liver maturation of breeder roosters using Ribo-Zero RNA sequencing. BMC Genomics 19(1):498. https://doi.org/10.1186/s12864-018-4891-7

Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, Jagannathan V et al (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res 45(8):1–12. https://doi.org/10.1093/nar/gkw1306

Xu X, Ji S, Li W, Yi B, Li H, Zhang H, Ma W (2017) LncRNA H19 promotes the differentiation of bovine skeletal muscle satellite cells by suppressing Sirt1/FoxO1. Cell Mol Biol Lett 22:10. https://doi.org/10.1186/s11658-017-0040-6

Xu E, Zhang L, Yang H, Shen L, Feng Y, Ren M, Xiao Y (2019) Transcriptome profiling of the liver among the prenatal and postnatal stages in chickens. Poult Sci 98(12):7030–7040. https://doi.org/10.3382/ps/pez434

Yang B, Jiao B, Ge W, Zhang X, Wang S, Zhao H, Wang X (2018) Transcriptome sequencing to detect the potential role of long non-coding RNAs in bovine mammary gland during the dry and lactation period. BMC Genomics 19(1):605. https://doi.org/10.1186/s12864-018-4974-5

Yang H, Ma J, Wang Z, Yao X, Zhao J, Zhao X, Wang F, Zhang Y (2020) Genome-wide analysis and function prediction of long noncoding RNAs in sheep pituitary gland associated with sexual maturation. Genes. https://doi.org/10.3390/genes11030320

Yin ZT, Lian L, Zhu F, Zhang Z-H, Hincke M, Yang N, Hou Z-C (2020) The transcriptome landscapes of ovary and three oviduct segments during chicken (*Gallus Gallus*) egg formation. Genomics 112(1):243–251. https://doi.org/10.1016/j.ygeno.2019.02.003

You Z, Zhang Q, Liu C, Song J, Yang N, Lian L (2019) Integrated analysis of LncRNA and MRNA repertoires in Marek's disease infected spleens identifies genes relevant to resistance. BMC Genomics 20(1):245. https://doi.org/10.1186/s12864-019-5625-1

Yue Y, Guo T, Liu J, Guo J, Yuan C, Feng R, Niu C, Sun X, Yang B (2015) Exploring differentially expressed genes and natural antisense transcripts in sheep (*Ovis Aries*) skin with different wool fiber diameters by digital gene expression profiling. PLoS ONE 10(6):e0129249. https://doi.org/10.1371/journal.pone.0129249

Zeng B, Chen T, Xie M-Y, Luo J-Y, He J-J, Xi Q-Y, Sun J-J, Zhang Y-L (2019) Exploration of long noncoding RNA in bovine milk exosomes and their stability during digestion in vitro. J Dairy Sci 102(8):6726–6737. https://doi.org/10.3168/jds.2019-16257

Zhang T, Zhang X, Han K, Zhang G, Wang J, Xie K, Xue Q, Fan X (2017) Analysis of long noncoding RNA and MRNA using RNA sequencing during the differentiation of intramuscular preadipocytes in chicken. PLoS ONE 12(2):e0172389. https://doi.org/10.1371/journal.pone.0172389

Zhang T, Zhang X, Han K, Zhang G, Wang J, Xie K, Xue Q (2017) Genome-wide analysis of LncRNA and MRNA expression during differentiation of abdominal preadipocytes in the chicken. G3 (Bethesda, Md.) 7(3):953–66. https://doi.org/10.1534/g3.116.037069

Zhang X, Chen M, Liu X, Zhang L, Ding X, Guo Y, Li X, Guo H (2020) A Novel LncRNA, Lnc 403, involved in bovine skeletal muscle myogenesis by mediating KRAS/Myf6. Gene

751(August):144706. https://doi.org/10.1016/j.gene.2020.144706

Zhang Z, Zhang S, Wang G, Feng S, Han K, Han L, Han L (2021) Role of MicroRNA and long non-coding RNA in Marek's disease tumorigenesis in chicken. Res Vet Sci 135(March):134–142. https://doi.org/10.1016/j.rvsc.2021.01.007

Zhao Yi, Li H, Fang S, Kang Y, Wei Wu, Hao Y, Li Z et al (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res 44(D1):D203–D208. https://doi.org/10.1093/nar/gkv1252

Zhao B, Chen Y, Shuaishuai Hu, Yang N, Wang M, Liu M, Li J, Xiao Y, Xinsheng Wu (2019) Systematic analysis of non-coding RNAs involved in the angora rabbit (*Oryctolagus Cuniculus*) hair follicle cycle by RNA sequencing. Front Genet 10:407. https://doi.org/10.3389/fgene.2019.00407

Zhao Z, Zou X, Tingting Lu, Deng M, Li Y, Guo Y, Sun B, Liu G, Liu D (2020) Identification of MRNAs and LncRNAs involved in the regulation of follicle development in goat. Front Genet 11:589076. https://doi.org/10.3389/fgene.2020.589076

Zheng J, Wang Z, Yang H, Yao X, Yang P, Ren C, Wang F, Zhang Y (2019) Pituitary transcriptomic study reveals the differential regulation of LncRNAs and MRNAs Related to prolificacy in different FecB genotyping sheep. Genes. https://doi.org/10.3390/genes10020157

Zhu S, Li W, Liu J, Chen C-H, Liao Qi, Ping Xu, Han Xu et al (2016) Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. Nat Biotechnol 34(12):1279–1286. https://doi.org/10.1038/nbt.3715

Zou C, Li J, Luo W, Li L, An Hu, Yuhua Fu, Hou Ye, Li C (2017a) Transcriptome analysis reveals long intergenic non-coding RNAs involved in skeletal muscle growth and development in pig. Sci Rep 7(1):8704. https://doi.org/10.1038/s41598-017-07998-9

Zou X, Wang J, Qu H, Lv XH, Shu DM, Wang Y, Ji J, He YH, Luo CL, Liu DW (2020) Comprehensive analysis of MiRNAs, LncRNAs, and MRNAs reveals Potential players of sexually dimorphic and Left-right asymmetry in chicken gonad during gonadal differentiation. Poult Sci 99(5):2696–2707. https://doi.org/10.1016/j.psj.2019.10.019

Zou C, Li S, Deng L, Guan Y, Chen D, Yuan X, Xia T, He X, Shan Y, Li C (2017b) Transcriptome analysis reveals long intergenic noncoding RNAs contributed to growth and meat quality differences between yorkshire and wannanhua pig. Genes. https://doi.org/10.3390/genes8080203

Zucchelli S, Fasolo F, Russo R, Cimatti L, Patrucco L, Takahashi H, Jones MH et al (2015) SINEUPs are modular antisense long noncoding RNAs that increase synthesis of target proteins in cells. Front Cell Neurosci 9(May):1–12. https://doi.org/10.3389/fncel.2015.00174

Review-2

Cytogenetic and Genome Research

Report

Cytogenet Genome Res 2022;162:405–527
DOI: 10.1159/000529376

Received: January 16, 2023
Accepted: January 22, 2023
Published online: January 30, 2023

Fourth Report on Chicken Genes and Chromosomes 2022

Prepared by

Jacqueline Smith[a] James M. Alfieri[b, c, d] Nick Anthony[e] Peter Arensburger[f] Giridhar N. Athrey[b, d] Jennifer Balacco[g] Adam Balic[a] Philippe Bardou[h] Paul Barela[i] Yves Bigot[j] Heath Blackmon[b, c] Pavel M. Borodin[k] Rachel Carroll[l] Meya C. Casono[i] Mathieu Charles[m] Hans Cheng[n] Maddie Chiodi[i] Lacey Cigan[i] Lyndon M. Coghill[o] Richard Crooijmans[p] Neelabja Das[i] Sean Davey[i] Asya Davidian[q] Fabien Degalez[r] Jack M. Dekkers[s, t] Martijn Derks[p] Abigail B. Diack[a] Appolinaire Djikeng[u] Yvonne Drechsler[v] Alexander Dyomin[q] Olivier Fedrigo[g] Steven R. Fiddaman[w] Giulio Formenti[g] Laurent A.F. Frantz[x, y] Janet E. Fulton[z] Elena Gaginskaya[q] Svetlana Galkina[q] Rodrigo A. Gallardo[s, A] Johannes Geibel[B, C] Almas A. Gheyas[a] Cyrill John P. Godinez[D] Ashton Goodell[i] Jennifer A.M. Graves[E, F] Darren K. Griffin[G] Bettina Haase[g] Jian-Lin Han[H, I] Olivier Hanotte[I, J, K] Lindsay J. Henderson[a] Zhuo-Cheng Hou[L] Kerstin Howe[M] Lan Huynh[N] Evans Ilatsia[O] Erich D. Jarvis[g] Sarah M. Johnson[i] Jim Kaufman[N, P, Q] Terra Kelly[s, A] Steve Kemp[R] Colin Kern[S] Jacob H. Keroack[i] Christophe Klopp[T] Sandrine Lagarrigue[r] Susan J. Lamont[s, t] Margaret Lange[U] Anika Lanke[V] Denis M. Larkin[W] Greger Larson[X] John King N. Layos[Y] Ophélie Lebrasseur[Z, α] Lyubov P. Malinovskaya[β] Rebecca J. Martin[Q] Maria Luisa Martin Cerezo[γ] Andrew S. Mason[δ] Fiona M. McCarthy[i] Michael J. McGrew[a, u] Jacquelyn Mountcastle[g] Christine Kamidi Muhonja[O, R] William Muir[ε] Kévin Muret[ζ] Terence D. Murphy[η] Ismael Ng'ang'a[x] Masahide Nishibori[θ] Rebecca E. O'Connor[G] Moses Ogugo[R] Ron Okimoto[e] Ochieng Ouko[O] Hardip R. Patel[ι] Francesco Perini[a, κ] María Ines Pigozzi[λ] Krista C. Potter[i] Peter D. Price[μ] Christian Reimer[B] Edward S. Rice[ν] Nicolas Rocos[N] Thea F. Rogers[ξ] Perot Saelao[s, S, o] Jens Schauer[B] Robert D. Schnabel[π] Valerie A. Schneider[η] Henner Simianer[C] Adrian Smith[w] Mark P. Stevens[a] Kyle Stiers[o] Christian Keambou Tiambo[R] Michele Tixier-Boichard[m] Anna A. Torgasheva[k] Alan Tracey[M] Clive A. Tregaskes[P, Q] Lonneke Vervelde[a] Ying Wang[s, S] Wesley C. Warren[ν, π] Paul D. Waters[ρ] David Webb[η] Steffen Weigend[B, C] Anna Wolc[t, z] Alison E. Wright[μ] Dominic Wright[γ] Zhou Wu[a] Masahito Yamagata[σ] Chentao Yang[τ] Zhong-Tao Yin[L] Michelle C. Young[i] Guojie Zhang[υ] Bingru Zhao[φ] Huaijun Zhou[s, S]

[a]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Edinburgh, UK; [b]Interdisciplinary Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, TX, USA; [c]Department of Biology, Texas A&M University, College Station, TX, USA; [d]Department of Poultry Science, Texas A&M University, College Station, TX, USA; [e]Cobb-Vantress, Siloam Springs, AR, USA; [f]Biological Sciences Department, California State Polytechnic University, Pomona, CA, USA; [g]The Rockefeller University, New York, NY, USA; [h]Université de Toulouse, INRAE, ENVT, GenPhySE, Sigenae, Castanet Tolosan, France; [i]University of Arizona, Tucson, AZ, USA; [j]PRC, UMR INRAE 0085, CNRS 7247, Centre INRAE Val de Loire, Nouzilly, France; [k]Department of Molecular Genetics, Cell Biology and Bioinformatics, Institute of Cytology and Genetics of Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia; [l]Department of Animal Sciences, Data Science and Informatics Institute, University of Missouri, Columbia, MO, USA; [m]University Paris-Saclay, INRAE, AgroParisTech, GABI, Sigenae, Jouy-en-Josas, France;

Karger@karger.com
www.karger.com/cgr

Correspondence to:
Jacqueline Smith, jacqueline.smith@roslin.ed.ac.uk

[n]USDA, ARS, USNPRC, Avian Disease and Oncology Laboratory, East Lansing, MI, USA; [o]Department of Veterinary Pathology, University of Missouri, Columbia, MO, USA; [p]Animal Breeding and Genomics, Wageningen University and Research, Wageningen, The Netherlands; [q]Saint Petersburg State University, Saint Petersburg, Russia; [r]INRAE, INSTITUT AGRO, PEGASE UMR 1348, Saint-Gilles, France; [s]Feed the Future Innovation Lab for Genomics to Improve Poultry, University of California, Davis, CA, USA; [t]Department of Animal Science, Iowa State University, Ames, IA, USA; [u]Centre for Tropical Livestock Genetics and Health (CTLGH) – The Roslin Institute, Edinburgh, UK; [v]College of Veterinary Medicine, Western University of Health Sciences, Pomona, CA, USA; [w]Department of Zoology, University of Oxford, Oxford, UK; [x]Queen Mary University of London, Bethnal Green, London, UK; [y]Palaeogenomics Group, Department of Veterinary Sciences, LMU Munich, Munich, Germany; [z]Hy-Line International, Research and Development, Dallas Center, IA, USA; [A]School of Veterinary Medicine, University of California, Davis, CA, USA; [B]Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, Neustadt, Germany; [C]Center for Integrated Breeding Research, University of Göttingen, Göttingen, Germany; [D]Department of Animal Science, College of Agriculture and Food Science, Visayas State University, Baybay City, Philippines; [E]Department of Environment and Genetics, La Trobe University, Melbourne, VIC, Australia; [F]Institute for Applied Ecology, University of Canberra, Canberra, ACT, Australia; [G]School of Biosciences, University of Kent, Canterbury, UK; [H]CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China; [I]International Livestock Research Institute (ILRI), Addis Ababa, Ethiopia; [J]Cells, Organisms and Molecular Genetics, School of Life Sciences, University of Nottingham, Nottingham, UK; [K]Centre for Tropical Livestock Genetics and Health, The Roslin Institute, Edinburgh, UK; [L]National Engineering Laboratory for Animal Breeding and Key Laboratory of Animal Genetics, Breeding and Reproduction, MARA, College of Animal Science and Technology, China Agricultural University, Beijing, China; [M]Wellcome Trust Sanger Institute, Hinxton, UK; [N]Institute for Immunology and Infection Research, University of Edinburgh, Edinburgh, UK; [O]Dairy Research Institute, Kenya Agricultural and Livestock Organization, Naivasha, Kenya; [P]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK; [Q]Department of Pathology, University of Cambridge, Cambridge, UK; [R]Centre for Tropical Livestock Genetics and Health (CTLGH) – ILRI, Nairobi, Kenya; [S]Department of Animal Science, University of California, Davis, CA, USA; [T]INRAE, MIAT UR875, Sigenae, Castanet Tolosan, France; [U]Department of Molecular Microbiology and Immunology, University of Missouri, Columbia, MO, USA; [V]BASIS Chandler High School, Chandler, AZ, USA; [W]Department of Comparative Biomedical Sciences, Royal Veterinary College, University of London, London, UK; [X]The Palaeogenomics and Bio-Archaeology Research Network, Research Laboratory for Archaeology and History of Art, The University of Oxford, Oxford, UK; [Y]College of Agriculture and Forestry, Capiz State University, Mambusao, Philippines; [Z]Centre d'Anthropobiologie et de Génomique de Toulouse (CAGT), CNRS UMR 5288, Université Toulouse III Paul Sabatier, Toulouse, France; [α]Instituto Nacional de Antropología y Pensamiento Latinoamericano, Ciudad Autónoma de Buenos Aires, Argentina; [β]Department of Cytology and Genetics, Novosibirsk State University, Novosibirsk, Russia; [γ]AVIAN Behavioural Genomics and Physiology, IFM Biology, Linköping University, Linköping, Sweden; [δ]Department of Biology, The University of York, York, UK; [ε]Department of Animal Sciences, Purdue University, West Lafayette, IN, USA; [ζ]Université Paris-Saclay, Commissariat à l'Energie Atomique et aux Energies Alternatives, Centre National de Recherche en Génomique Humaine, Evry, France; [η]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; [θ]Laboratory of Animal Genetics, Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima, Japan; [ι]The John Curtin School of Medical Research, Australian National University, Canberra, ACT, Australia; [κ]Department of Agricultural, Food and Environmental Sciences, University of Perugia, Perugia, Italy; [λ]INBIOMED (CONICET-UBA), Facultad de Medicina, Universidad de Buenos Aires, Buenos Aires, Argentina; [μ]Ecology and Evolutionary Biology, School of Biosciences, University of Sheffield, Sheffield, UK; [ν]Department of Animal Sciences, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA; [ξ]Department of Molecular Evolution and Development, University of Vienna, Vienna, Austria; [ο]Veterinary Pest Genetics Research Unit, USDA, Kerrville, TX, USA; [π]Department of Animal Sciences, University of Missouri, Columbia, MO, USA; [ρ]School of Biotechnology and Biomolecular Science, Faculty of Science, UNSW Sydney, Sydney, NSW, Australia; [σ]Center for Brain Science, Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA; [τ]BGI-Shenzhen, Shenzhen, China; [υ]Center for Evolutionary and Organismal Biology, Zhejiang University School of Medicine, Hangzhou, China; [φ]College of Animal Science and Technology, Nanjing Agricultural University, Nanjing, China

Introduction

(Prepared by J. Smith)

The chicken continues to hold its position as a leading model organism within many areas of research, as well as being a major source of protein for human consumption. The First Report on Chicken Genes and Chromosomes [Schmid et al., 2000], which was published in 2000, was the brainchild of the late, and sadly missed, Prof. Michael Schmid of the University of Würzburg. It was a publication bringing together updates on the latest research and resources in chicken genomics and cytogenetics. The success of this first report led to the subsequent publication of the Second [Schmid et al., 2005] and Third Report on Chicken Genes and Chromosomes [Schmid et al., 2015] – each also proving popular references for the research community. It is now our pleasure to be able to introduce publication of the Fourth Report. Being 7 years since the last report, this publication captures the many advances that have taken place during that time. This includes pre-

sentation of the detailed genomic resources that are now available, largely due to increasing capabilities of sequencing technologies and which herald the pangenomic age, allowing for a much richer and more complete knowledge of the avian genome. Ongoing cytogenetic work also allows for examination of chromosomes, specific elements within chromosomes, and the evolutionary history and comparison of karyotypes. We also examine chicken research efforts with a much more "global" outlook with a greater impact on food security and the impact of climate change, and highlight the efforts of international consortia, such as the Chicken Diversity Consortium. We dedicate this Report to Michael.

Evolution of Protein-Coding and Long Noncoding Genes of the Chicken Genome through the Different Genome Assemblies and Their Associated Annotations

(Prepared by F. Degalez, K. Muret, and S. Lagarrigue)

The chicken genome was the first avian genome sequenced because of its importance in human food production, in fundamental biology like the study of devel-

opment or gene function conservation across evolution [International Chicken Genome Sequencing Consortium, 2004]. Since its first version in February 2004 (Galgal2/WUGSC1.0), five new genome assemblies have been released, each improving the genome sequences' accuracy. Along with these genome assemblies, numerous genome annotations were released, providing at least models for gene loci and transcripts supporting them. Since the first annotated version (Ensembl v22 - May 25, 2004) associated with the Galgal2 assembly, the number of genes and the diversity of their biotypes have increased, especially in 2015 with the introduction of long noncoding RNAs (lncRNAs), which is concurrent with the first initiatives of lncRNA annotation [Chodroff et al., 2010; Necsulea et al., 2014; Li et al., 2015; Muret et al., 2017].

LncRNAs represent a large and heterogeneous class of genes defined by transcripts longer than 200 nucleotides without coding-potential capabilities [Derrien et al., 2012]. They represent a variety of regulatory elements implied in gene expression and can act at different levels by using diverse biological mechanisms based on DNA, RNA, or protein interactions [Guh et al., 2020]. As illustrated in Figure 10, lncRNAs can interact with DNA, RNA, and proteins and act at different molecular levels: nuclear organization (e.g., MALAT1 [Wang X et al., 2021b]/NEAT1 [Yamazaki et al., 2018]) (Fig. 10A), genome integrity (e.g., TERRA [Barral and Déjardin, 2020]) (Fig. 10B), histone marks modification for silencing (e.g., Fendrr [Grote et al., 2013]) or activating (e.g., GATA3-AS1 [Gibbons et al., 2018]) gene transcription (Fig. 10C), loop formation to connect enhancers to promoter regions (e.g., MYMLR [Kajino et al., 2019]) (Fig. 10D). LncRNAs can modulate RNA splicing (e.g., linc-HELLP [van Dijk et al., 2015]) (Fig. 10E), miRNA maturation (e.g., CCAT2 [Yu et al., 2017]/uc.372 [Guo et al., 2018]) (Fig. 10F), and protein translation (e.g., BC1 [Wang et al., 2002]/MCM3AP-AS1 [Guo C et al., 2020]) (Fig. 10I) or their activity (e.g., NORAD [Munschauer et al., 2018]) (Fig. 10K). They can also control the stabilization or the degradation of molecules as miRNAs (e.g., ROR [Li C et al., 2017]/DSCR8 [Wang Y et al., 2018b]) (Fig. 10G), mRNAs (e.g., PTB-AS [Zhu L et al., 2019]/TINCR [Xu et al., 2015]) (Fig. 10H), and proteins (e.g., PiHL [Deng et al., 2020]/MALAT1 [Yan et al., 2016]) (Fig. 10J). LncRNAs can host small ORFs [Choi et al., 2019] which code for peptides (e.g., CASIMO1 [Polycarpou-Schwarz et al., 2018]/DWORF [Nelson et al., 2016]) (Fig. 10M) or host in their introns small RNAs [Sun Q et al., 2021] (e.g., MCM7 [Agranat-Tamir et al., 2014]/DLEU2 [Morenos et al., 2014]) (Fig. 10N). They can control protein transfers from cytoplasm to nucleus (e.g., NRON [Willingham et al., 2005]) or from nucleus to cytoplasm (e.g., Discn [Wang L et al., 2021]) (Fig. 10L). Finally, they can migrate to other cells with exosomes (e.g., ZFAS1 [Pan et al., 2017]/GAS5 [Chen et al., 2017]) (Fig. 10O).

Through their key roles in gene regulation, lncRNAs are consequently involved in diverse biological and pathophysiological processes [Ponting et al., 2009; Muret et al., 2019; Gil and Ulitsky, 2020; Statello et al., 2021]. Moreover, since most of the trait-associated variations identified by genome-wide association studies (GWAS) concerned noncoding intervals of the genome [Manolio et al., 2009; Bouwman et al., 2018], this reinforces the need to characterize the regulatory regions of domesticated species such as lncRNA genes. LncRNA genes have different characteristics compared to protein-coding genes (PCGs). They are less expressed [Derrien et al., 2012; Muret et al., 2017; Le Béguec et al., 2018; Jehl et al., 2020], explaining why they have been detected only recently – i.e., this last decade – by high-throughput transcriptome sequencing technologies (RNA-seq). Furthermore, lncRNA expression is more specific to tissues, life stages, and conditions than that of PCGs [Cabili et al., 2011; Derrien et al., 2012; Jehl et al., 2020]. The identification of these genic entities is therefore dependent on the variety of RNA-seq data available to detect them.

After presenting the different chicken genome assemblies developed over the last 2 decades, we discuss the associated genome annotations provided by NCBI's RefSeq and EMBL-EBI's Ensembl, the two reference annotation databases. We characterize them in terms of number of gene and transcript models, variety of biotypes, or in terms of models that are shared by the two reference databases. We show that lncRNA loci are even less well-known than PCG ones, although, for the latter, knowledge of their transcripts can be further improved. Finally, we discuss the impacts of these weaknesses and the value of gathering different genome annotation resources, in particular, for a better description of lncRNA loci, and then present two initiatives. The MANE project yet limited to the human genome aims to synergize the NCBI's RefSeq and EMBL-EBI's Ensembl "gene" databases to establish a consensus annotation. The second project, specifically realized for the chicken, is to provide a "gene" database built from various resources including the NCBI's RefSeq and EMBL-EBI's Ensembl databases and other resources such as FAANG multi-tissue resources and NONCODE database. This gene catalog is maintained at each significant update in chicken genome assembly and

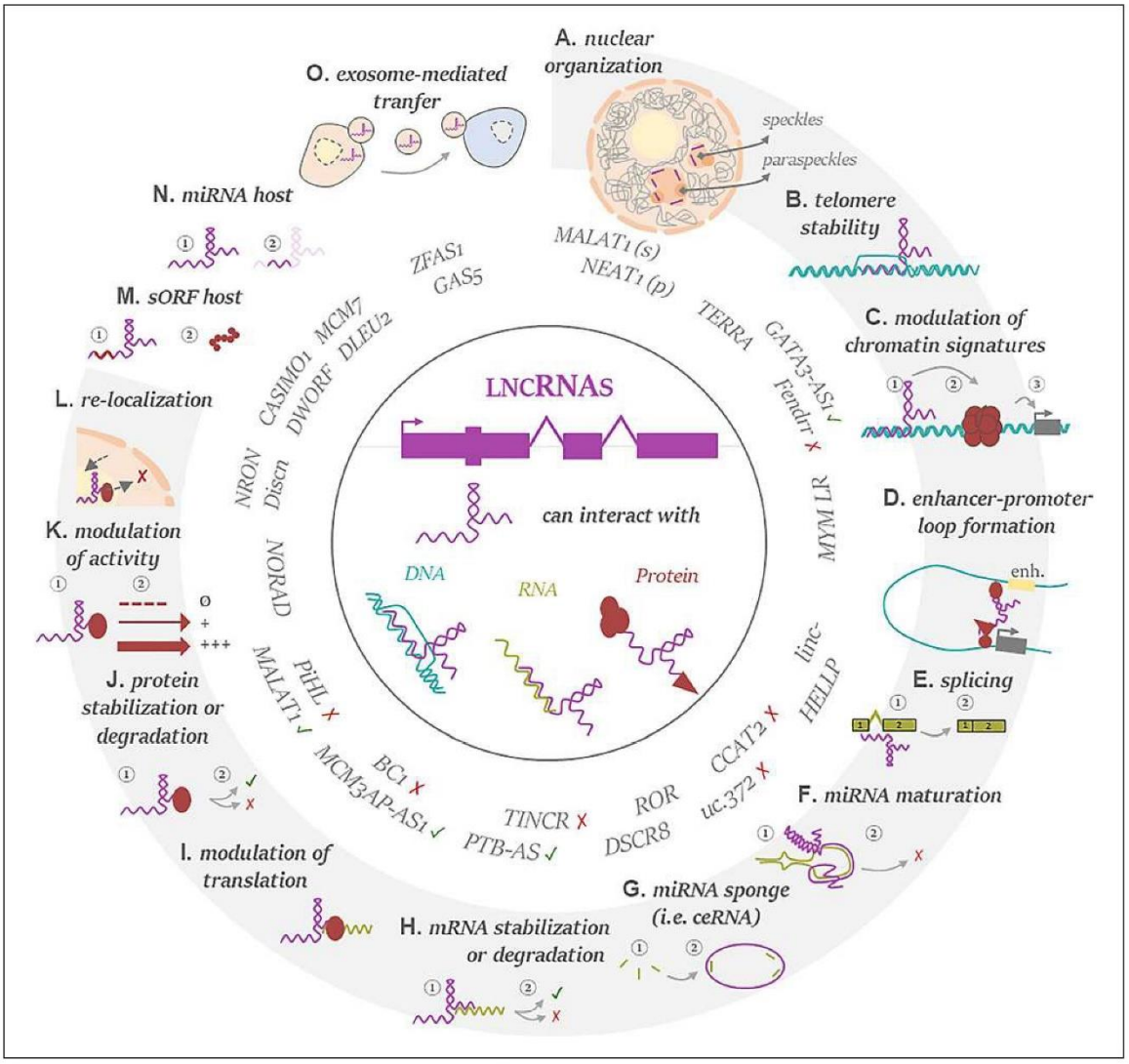


Fig. 10. Different mechanisms of lncRNA roles. Effects at the nuclear and telomere (**A**, **B**), transcriptional (**C**, **D**), post-transcriptional (**E–H**), translational (**I**), and post-translational levels (**J–L**). Role as small ORF host (**M**) and small noncoding RNA host (**N**). Implication in the exosome-mediated transfer (**O**). In purple, lncRNA; in blue, DNA; in green, other RNAs; in dark red, proteins. For more examples, other genes are presented in Muret et al. [2019], specifically genes involved in the regulation of lipid metabolism and their regulatory mechanisms.

genome annotation, the last version of June 2022, associated to the GRCg7b assembly, being composed of 23,926 PCGs and 44,428 lncRNA genes (available at http://www.fragencode.org).

Evolution of the Reference Sequence of the Chicken Genome

As illustrated in Figure 11a, while the overall coverage of Galgal2/WUGSC1.0 was 6.63× [International Chicken Genome Sequencing Consortium, 2004], this parameter is

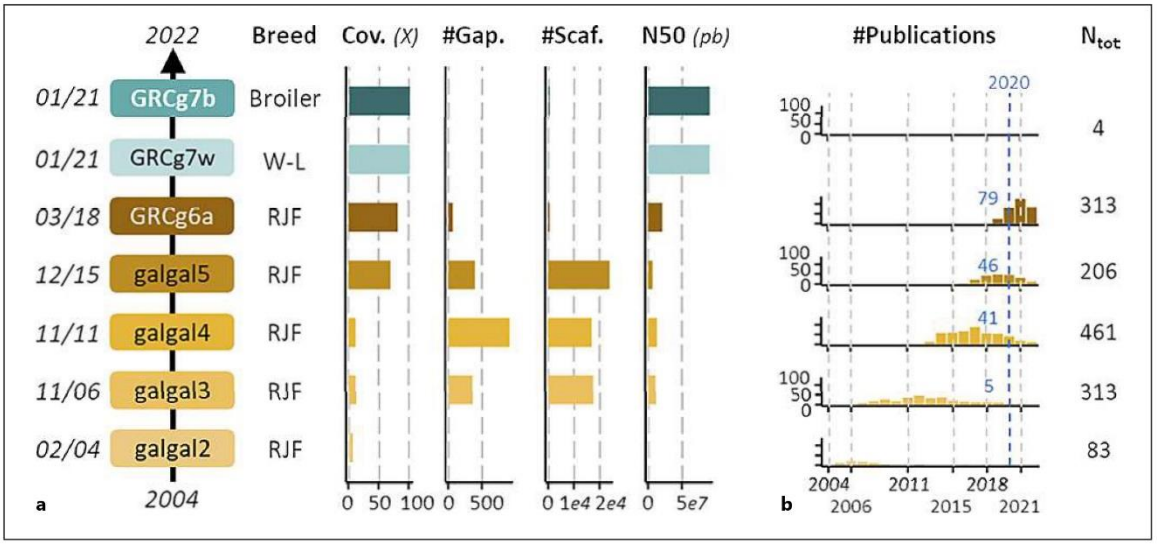


Fig. 11. Assembly versions associated with the chicken genome (**a**) and the number of publications associated with them (**b**). **a** RJF, Red Jungle fowl; W–L, White Leghorn; Cov., coverage; Scaf, scaffold; N50, scaffold N50. For more details, see online supplementary Material 5, Table 1. **b** Blue numbers, articles published using the corresponding assembly during the year 2020. Identification was made on PubMed Central by searching for the assembly name in different formats (e.g., GRCg6a or GRCg6 or Galgal6).

doubled for the two successive assemblies, i.e., Galgal3/WUGSC2.1 [NCBI RefSeq, 2006] and Galgal4 [NCBI RefSeq, 2011], which were released in November 2006 and 2011, respectively. Compared to the first version, these include 33 chromosomes (1–28; 32; W/Z; mitochondrial chromosome [MT]) but the number of scaffolds remained very high (~17,000), with 915 gaps between the scaffolds and a scaffold N50 which was quite low (~12 Mb) showing the incompleteness of the chicken genome sequence. Note that scaffold N50 is defined as the sequence length of the shortest scaffold at 50% of the total genome length. With the release of the Galgal5 assembly in December 2015 and with the improvement of sequencing technologies [NCBI RefSeq, 2015; Warren et al., 2017], the average coverage exploded and reached a global depth of 70×, leading to a better knowledge of the genome sequence. A few chromosomes were newly defined (addition of chromosomes 29–31). However, the quality of this genome sequence remained low with a high number of scaffolds (~24,000) and a lower scaffold N50 (~6.4 Mb) than before. These weak performances are likely due to long-read sequencing, which improved the detection of smaller scaffolds thus decreasing the N50 value [Warren et al., 2017].

In March 2018, a new assembly called GRCg6a was released by the Genome Reference Consortium, which has taken the lead concerning the chicken genome assembly previously managed by the International Chicken Genome Consortium [NCBI RefSeq, 2018]. Tremendous progress — due to the addition of long read sequences, improved de novo assembly algorithms, manual annotation of contigs, and integration of finished BAC clone sequences — was made regarding the genome accuracy, as indicated by the drop in the number of scaffolds (from ~24,000 to ~500) with only 68 gaps between the scaffolds and an increase in the scaffold N50 (from ~6.4 M to ~20 M).

Furthermore, with the latest GRCg7 genome assemblies [NCBI RefSeq, 2021a, b], the knowledge of the chicken genome sequence improved even more in two main ways. First, the accuracy of the genome sequence increased due to improvements in sequencing and especially assembly technologies. The chicken genome is now composed of 42 chromosomes (1–39; W/Z; MT), reaching the number observed in the chicken karyotype. This assembly includes more microchromosomes with ~250 scaffolds with no gap between scaffolds, and the scaffold N50 reaching 90 Mb. Second, whereas a Red Jungle Fowl breed (known as RJF #256) was always used in previous assembly versions, a trio of chickens from diverse breeds was used for GRCg7. The new reference sequence was

generated from a female offspring from a cross between a broiler female and a white leghorn laying male. The RJF breed, considered to be the descendant of domestic chickens, was used as a good representation of broiler and layer chicken breeds; however, such a choice has a significant impact on the detection of variants leading to the identification of false positives. Since actual breeds have diverged from the RJF, a variation (e.g., SNP) at one position may be detected according to the RJF genome sequence, whereas this position was fixed in the population of interest. As an illustration, we have previously shown, using RNA-seq data from the liver of 11 different breeds (~750 RNA-seq of ~400 birds) aligned on the Galgal5 assembly, that the SNP number with reliable genotypes was on average 549,634 per population, but this number dropped to 339,539 (−38.2%) with a minor allele frequency ≥10% [Jehl et al., 2021]. This drop is mainly due to fixed variants in the populations since the number decreased to 438,837 (−20.2%) after only excluding the fixed variations.

Consequently, two genome assemblies were released in January 2021: GRCg7b representing the broiler breed and considered as the new genome reference, and GRCg7w representing the laying breed and considered as an alternative.

Because of the quite frequent change of the genome assembly compared to the time needed to conduct and publish a scientific study, a lot of works are published in outdated versions which can lead to the publication of misleading results and in disagreement with more recent versions (Fig. 11b). For example, in 2020, two years after the release of GRCg6a, 79 studies using this genome reference were published against 46, 41, and 5 published with the Galgal5, Galgal4, and Galgal3 versions, respectively. Some tools such as Liftoff [Shumate and Salzberg, 2020] or LiftOver [Kuhn et al., 2013] can be used to convert coordinates from one version to another. The first tool is based on the alignments of the gene features from one annotation to another, whereas the second tool is based on alignments of the best/longest syntenic regions for each region of the genome between assemblies (chain files). However, the use of these tools must be done with caution, especially for remote versions, because of important changes in the genome sequence.

Evolution of the Two Reference Genome Annotations: A Breakthrough in 2015 with the Apparition of the lncRNA Gene Biotype

Genome annotation is not only evolving according to the version of the genome assembly but also to the evolution of annotation bioinformatics pipelines and data resources, mainly composed nowadays of RNA-seq data. In Figure 12a, genome annotations, from 2004 with the Galgal2/WUGSC1.0 assembly through 2022 with the GRC7b assembly, produced by the reference centers, NCBI's RefSeq and EMBL-EBI's Ensembl, have been analyzed. As illustrated, the gene number has increased, particularly due to the apparition in 2015 of lncRNAs, with 5,763 and 4,641 lncRNAs modeled by NCBI's RefSeq (v103) and EMBL-EBI's Ensembl (v94). This increase continues with the last genome annotation v107 (associated with the GRCg7b assembly) provided by EMBL-EBI's Ensembl with 11,944 lncRNAs compared with 5,504 for v106 (associated with GRCg6a). The number of PCGs remains constant at around 17,000 (see further for more explanation regarding such evolution).

In parallel to the gene number, it is important to make some comments about the transcript models that support these genes. As observed in Figure 12b, the transcript models can still be improved, as illustrated by the numerous changes observed between the two versions v105 and v106 of NCBI's RefSeq [NCBI RefSeq, 2022]. Only 6.4% of 93,980 transcripts identified in the 106 version are identical to those found in version v105. Such results can also be observed between genome annotations of different genome assemblies (not shown), or between genome annotations from the two reference centers, NCBI and EMBL-EBI, as illustrated in the next section.

Differences between the Latest NCBI RefSeq and EMBL-EBI Ensembl Genome Annotations

For the same genome assembly, the two genome annotation bioinformatics centers, EMBL-EBI and NCBI, do not provide the same annotations, as illustrated in Figure 13. First, EMBL-EBI's Ensembl provides twice the number of lncRNA gene models compared to NCBI's RefSeq (shown in Fig. 13a) resulting in a total of 30,108 gene models (associated to 72,689 transcripts) including 17,007 PCGs, 11,944 lncRNAs, and 674 miRNAs compared with 25,638 gene models (associated with 85,704 transcripts) including 18,024 PCGs, 5,791 lncRNAs, and 799 miRNAs for RefSeq. These differences can be explained by the sample datasets and the annotation pipeline thresholds used specifically by the two bioinformatics centers. For example, NCBI's RefSeq does not consider lncRNAs supported by a single mono-exonic transcript in contrast to EMBL-EBI's Ensembl (with 1,157 lncRNA loci). Second, using the "GffCompare" software [Pertea and Pertea, 2020], we observed that most of the transcript models are different between the two ge-

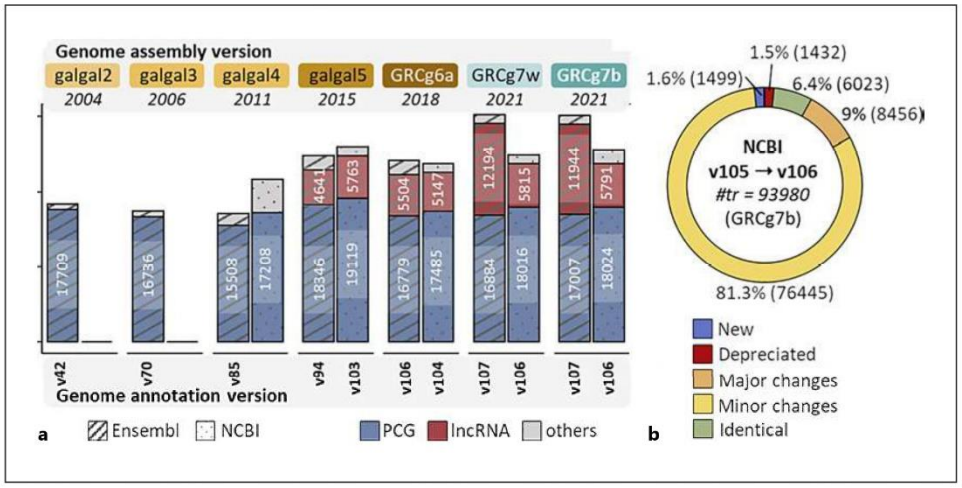


Fig. 12. Gene numbers provided by NCBI's RefSeq and EMBL-EBI's Ensembl according to the genome annotation and genome assembly versions (**a**) and transcript model changes between two genome annotation versions from NCBI RefSeq for the same assembly – GRCg7b (**b**). **a** PCG, protein-coding gene; lncRNA: long noncoding RNA. **b** Comparison between versions 105 and 106 provided by NCBI [NCBI RefSeq, 2022]. Briefly, a score (between 0 and 1) for current and previous transcript features is calculated based on overlap in exon sequence and matches in exon boundaries. Pairs of current and previous features were categorized based on these scores and considering changes in attributes. New, new transcript models; Deprecated, transcripts removed or merged in the new version; Major changes, changes with great impact on the sequence or on the transcript attributes; Minor changes, minimal change ensuring similarity.

nome annotations, as shown in Figure 13b. Among the 72,579 transcripts from EMBL-EBI's Ensembl considered in the analysis, only 17.8% are strictly equal in the NCBI's RefSeq annotation. More than half (55.9%) are identified as new isoforms of an existing locus and 26.1% (18,922) transcripts are associated with 9,958 new gene loci resulting in more than one-third of the 30,108 gene models from EMBL-EBI's Ensembl not being known in NCBI's RefSeq.

Important differences exist between PCG and lncRNA transcripts. For PCG, most of transcripts from EMBL-EBI's Ensembl (70.8%) are new isoforms of the same gene loci existing in the two databases. These results show that the transcript isoforms are not well described with current RNA-seq resources. Indeed, most of RNA-seq data available in the public database are short-read RNA-seq; the long-read RNA-seq studies using the new technologies such as ONT or PacBio are still very limited [Kuo et al., 2017; Guan et al., 2022] due to the cost of these technologies and their low sequencing depth. For lncRNA, most of the lncRNA transcripts from EMBL-EBI's Ensembl (77.4%) are considered as new loci compared to NCBI's RefSeq. The main cause of this very low gene overlap between the two genome annotations is the difficulty in capturing and therefore modeling lncRNAs compared to PCGs, due to specific features of lncRNA. First, lncRNAs are characterized by a global low expression; around less than 10% of the total reads of a sample analyzed by common technologies support lncRNA transcripts [Lagarrigue et al., 2022]. Second, they are tissue-, developmental stage-, and condition-specific [Cabili et al., 2011; Derrien et al., 2012; Jehl et al., 2020], conditions which are not covered by the limited number of RNA-seq samples used by the reference genome annotation centers compared to the tens of thousands of short-read RNA-seq generated by the avian scientific community which are available in the public database.

Moreover, the transcript models from NCBI's RefSeq are significantly longer than those of EMBL-EBI's Ensembl, as shown in Figure 13c, particularly for lncRNAs (almost twice the length), with nearly two supplementary exons by transcript (resp. a median of 5 vs. 10 exons/transcript, $p < 10^{-16}$) for both PCG and lncRNA models, with median exon sizes which remain similar (~250 bp). Moreover, NCBI's RefSeq provides a higher extreme distribution of transcripts per gene for PCGs compared to

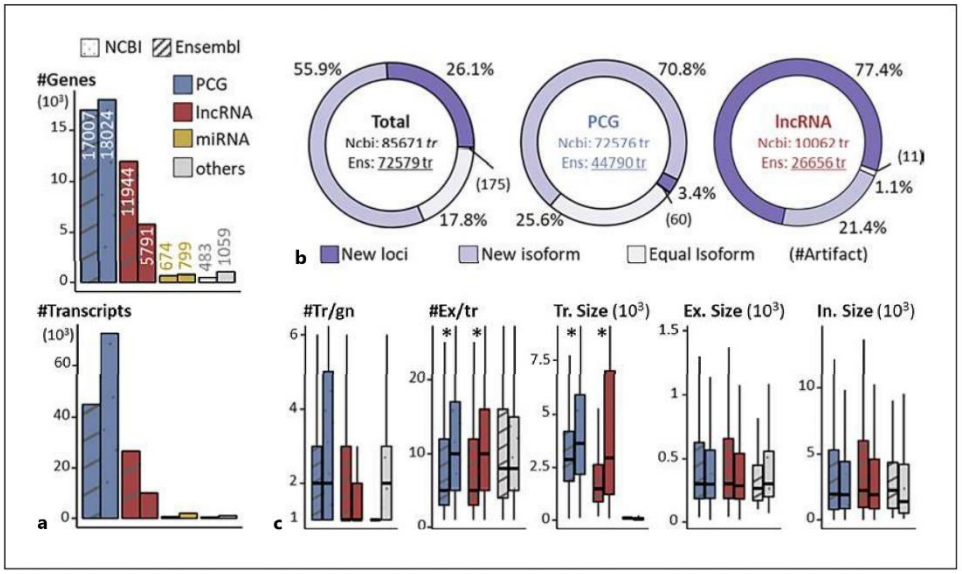


Fig. 13. Features of the current NCBI RefSeq (v106) and EMBL-EBI Ensembl (v107) genome annotations based on the latest GRCg7b genome assembly. **a** Number of genes and transcripts according to gene biotypes for the two genome annotations. **b** The transcript models were compared between the two annotations according to 4 main classes (Equal isoform, New isoform, New loci, and Artifacts) according to the software "GffCompare" (options: -S --no-merge) [Pertea and Pertea, 2020]. **c** #Tr/gn, number of transcripts per gene; #Ex/tr, number of exons per transcript; Tr. Size, transcript size considering only exonic regions; Ex. Size, exon size; In. Size, intron size. The median transcript sizes between RefSeq and Ensembl are 3,465 bp versus 2,317 bp, respectively, $p < 10^{-16}$ (Wilcoxon rank sum test); for PCG 3,634 bp versus 2,870 bp, * $p < 10^{-16}$; for lncRNAs 2,952 bp versus 1,487 bp, * $p < 10^{-16}$.

EMBL-EBI's Ensembl (resp. 5 vs. 3 for the third quartile and resp. 9 and 5 for the last decile). Note that these numbers are far below what is described in human (resp. 6, 11, 18 transcripts per gene for the median, the third quartile, and the last decile (EMBL-EBI's Ensembl v107 with the GRCh38.13 assembly). This discrepancy can be explained in part by the variety of samples each reference used. EMBL-EBI's Ensembl combines a short-read RNA-seq dataset of 21 tissues from the Roslin Institute (1 stage-condition of a same breed per tissue, 21 samples of individual pool) with a short-read dataset of 7 tissues from the GENE-SWitCH project (3 stages of a same breed, 84 samples) and a long-read dataset of 6 tissues (7 samples) [EMBL EBI's Ensembl, 2022]. NCBI's RefSeq integrates data from various projects representing more than 20 tissues, different development stages and breeds for a total of 100 and 89 samples for short-read and long-read RNA-seq, respectively, in addition to "Cap Analysis Gene Expression" (CAGE) data including those from the FANTOM project [Lizio et al., 2017] for improving the annotation of transcription start sites [NCBI RefSeq, 2022]. For lncRNAs, the pattern in the distribution of the number of transcripts per gene is inverted between NCBI's RefSeq and EMBL-EBI's Ensembl with respectively 2 versus 3 for the third quartile and 3 and 5 for the last decile. Interestingly, these numbers are of the same order of magnitude in human (resp. 1, 2, and 5 transcripts per gene for the median, the third quartile, and the last decile), highlighting the general difficulty in capturing the transcript models associated with lncRNA genes.

Interest in an Annotation Combining NCBI's RefSeq and EMBL-EBI's Ensembl

In summary, different genome annotations coexist with important differences in transcript models for PCGs and gene models for lncRNAs. Initiatives like the MANE project [Morales et al., 2022] for the human genome aim to synergize the NCBI's RefSeq and EMBL-EBI's Ensembl reference genome annotations to establish a consensus, although, so far, these efforts have focused only on PCGs. Such initiatives have yet to exist for livestock species, especially chicken. So far, most RNA-seq studies have ana-

lyzed gene expression and focus only on PCGs, using only one of these two reference annotations. As previously reported, the last two chicken reference genome annotations are quite similar in terms of PCG loci. Indeed, 18,024 and 17,007 PCG loci are respectively annotated for NCBI's RefSeq and EMBL-EBI's Ensembl; 15,711 (87.2%) loci from NCBI's RefSeq are shared with 15,848 (93.8%) loci from EMBL-EBI's Ensembl, even if most of the transcript models supporting these PCGs are different. However, these numbers drop for the 5,791 and 11,944 lncRNA loci respectively from NCBI's RefSeq and EMBL-EBI's Ensembl where 2,008 (34.7%) loci from NCBI's RefSeq are shared with only 2,118 (17.7%) loci from EMBL-EBI's Ensembl. Therefore, the use of only one of the two reference annotations enables the investigation of most PCG loci but can bias the study of lncRNA loci. Moreover, even when the expression is quantified at the gene level and not the transcript level, the high difference of transcript models previously reported — even for PCG loci — can have an impact. Thus, in the context of gene expression studies, results could differ depending on the annotation used [Zhao and Zhang, 2015]. Furthermore, the difference between transcript models, especially for PCGs, may have an important impact on variant prediction [McCarthy et al., 2014].

Concerning the recent studies interested in lncRNA gene expression, most of them have not used a reference genome annotation because of the very limited number of lncRNA loci represented in the versions – before the latest EMBL-EBI's Ensembl v107 – and produce a de novo annotation from investigators' own samples [for review, see Lagarrigue et al., 2022]. Such an approach is due to the recent democratization of RNA-seq data and the RNA-seq processing, gene modeling, and lncRNA prediction pipelines. However, these genome annotations are specific to one tissue or a set of tissues and characterized by their own gene identifiers, making result comparison difficult from one study to another. As reported in recent reviews [Kosinska-Selbi et al., 2020; Lagarrigue et al., 2022], the number of such publications has been constantly growing since 2015, with most of them focusing on the tissue-specific expression of lncRNAs or their differential expression in a given tissue between breeds or animal groups contrasted for an economically important trait in the species of interest. In most of these studies, a few lncRNAs have been highlighted as associated with the trait or tissue of interest whereas the lncRNA catalogues are not really exploited by the scientific community.

In parallel to these tissue-specific studies, a few multi-tissue studies have been performed in order to provide a more comprehensive annotation of lncRNAs and considering their high tissue specificity. We can point two studies, that are part of the Functional Annotation of ANimal Genome consortium (FAANG) [Andersson et al., 2015], which have provided a multispecies lncRNA annotation: the first, Foissac et al. [2019], used 3 tissues of 4 female and male biological replicates of 4 farm species including the chicken updated to 12 tissues (personal communication); the second one, Kern et al. [2018], used 8 tissues of 2 female and male biological replicates of chicken, pig and cattle and was recently updated to 19 tissues for the chicken species [Guan et al., 2022]. Nevertheless, for a given species such as chicken, these studies remain limited due to the range of tissues, stage of development, condition that may exist.

In this context, we proposed since 2020 to provide a comprehensive gene catalog for chicken by gathering different resources, including EMBL-EBI's Ensembl, NCBI's RefSeq, and other multi-tissue databases, that we update at each important change of chicken genome assembly and annotation. Since the release of the new GRCg7b chicken genome sequence, we have recently updated the gene catalogue of 52,075 genes published in 2020 [Jehl et al., 2020], considering the last NCBI's RefSeq and EMBL-EBI's Ensembl annotations available in June 2022. First, we gathered the two genome annotation references, i.e., the v106 of RefSeq and the v107 of Ensembl resources. In addition to these two references, we chose to gather the two updated FAANG multi-tissue resources described above [Foissac et al., 2019; Guan et al., 2022], in which lncRNAs have been modeled in parallel with the PCG loci. The NONCODE resource composed only of lncRNA loci has also been used, even if this resource has not been updated since 2014 for the chicken [Zhao et al., 2021]. As a result, the EMBL-EBI's Ensembl and NCBI's RefSeq references grew respectively from 17,007 to 18,024 to 23,926 PCGs and from 11,944 to 5,791 to 44,428 lncRNA genes. This atlas associated to GRCg7b assembly is publicly available at http://www.fragencode.org [Degalez et al., in preparation] as the previous ones published in 2020 and associated to the Galgal5 and GRCg6a genome assemblies. In addition to the gene atlas (i.e., gtf file), a functional annotation of the genes across 40 tissues using different public resources is also provided as well as the lncRNA gene naming according to the official HUGO gene nomenclature committee (HGNC). Briefly, for the lncRNAs with an unknown function (frequent cases), the lncRNA adopts the symbol gene name of the gene harboring it, enriched by a suffix describing its genom-

ic location. For more information on the lncRNA nomenclature, see Wright [2014] and Muret et al. [2019] (online suppl. Material 5).

Conclusion

This review provides an overview of the evolution of chicken genome assemblies from 2004 to June 2022 and their genome annotations provided by the two most widely used annotation databases, NCBI's RefSeq and EMBL-EBI's Ensembl. We show a great evolution of the genome assembly through 6 different versions due to various technical and technological advances, the latest GRCg7b offers a genome reference sequence composed of 42 chromosomes (1–39; W/Z; MT), reaching the number observed in the chicken karyotype with more microchromosomes than the previous versions and with no gap between the ~250 scaffolds. Moreover, we show that the annotation of the chicken genome is constantly evolving according to the version of the genome assembly, the evolution of bioinformatic annotation pipelines, and the RNA-seq data resources. We can highlight the recent emergence, in 2015, of lncRNA models in genome annotations associated with the Galgal5 genome assembly. Concerning the last GRCg7b genome assembly, the two reference genome annotations are quite different with 18,024 PCGs and 5,791 lncRNAs reported for NCBI's RefSeq and 17,007 PCGs and 11,944 lncRNAs for EMBL-EBI's Ensembl. The PCG entities mainly differ at the transcript model level whereas lncRNAs differ both at the transcript and gene loci levels. Gene loci display a very low overlap mainly explained by the specific features of lncRNAs (low expression, high tissue-, condition-specificity, …) and the limited number of RNA-seq samples used for generating these catalogs. To facilitate the reconstruction of full-length transcript models, and so accurate gene models, annotation centers will benefit in the near future from new technologies such as ONT or PacBio allowing long-read RNA sequencing. However, for properly catching lncRNAs, the low sequencing depths of these long-read technologies compared to short-read RNA-seq require preliminarily capture strategies used to boost the concentration of low-abundance transcripts in cDNA libraries. Such strategies have been applied to 4 human and mouse tissues by the GENCODE consortium [Lagarde et al., 2017]. However, the low sequencing depths and the high cost of these technologies limit for the moment their wider use. The main fuel of the genome annotation databases remains the short-read RNA-seq massively generated by the scientific community. In this context, to increase the completeness of the chicken genome annotation, especially lncRNAs, we highlight the interest to combine the two reference NCBI's RefSeq and EMBL-EBI's Ensembl genome annotation databases and even other databases and present two initiatives. One of them, applied to the chicken species and updated at each important change of the reference annotation, provides a catalogue of 23,926 PCG and 44,428 lncRNA gene models which includes all the gene loci of the last versions (June 2022) of NCBI's RefSeq and EMBL-EBI's Ensembl.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

F. Degalez is a Ph.D. fellow supported by the Brittany regio (France) and the INRAe's Animal Genetics division.

References

Agranat-Tamir L, Shomron N, Sperling J, Sperling R. Interplay between pre-mRNA splicing and microRNA biogenesis within the supraspliceosome. Nucleic Acids Res. 2014 Apr;42(7):4640–51.

Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biology. 2015 Mar 25;16(1):57.

Barral A, Déjardin J. Telomeric Chromatin and TERRA. Journal of Molecular Biology. 2020 Jul 10;432(15):4244–56.

Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 2018 Mar;50(3):362–7.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011 Sep 15;25(18):1915–27.

Chen L, Yang W, Guo Y, Chen W, Zheng P, Zeng J, et al. Exosomal lncRNA GAS5 regulates the apoptosis of macrophages and vascular endothelial cells in atherosclerosis. PLOS ONE. 2017 Sep 25;12(9):e0185406.

Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, et al. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol. 2010;11(7):R72.

Choi SW, Kim HW, Nam JW. The small peptide world in long noncoding RNAs. Brief Bioinform. 2019 Jun 3;20(5):1853–64.

Deng X, Li S, Kong F, Ruan H, Xu X, Zhang X, et al. Long noncoding RNA PiHL regulates p53 protein stability through GRWD1/RPL11/MDM2 axis in colorectal cancer. Theranostics. 2020 Jan 1;10(1):265–80.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012 Jan 9;22(9):1775–89.

van Dijk M, Visser A, Buabeng KML, Poutsma A, van der Schors RC, Oudejans CBM. Mutations within the LINC-HELLP non-coding RNA differentially bind ribosomal and RNA splicing complexes and negatively affect trophoblast differentiation. Hum Mol Genet. 2015 Oct 1;24(19):5475–85.

EMBL EBI's Ensembl. Gallus_gallus - Ensembl genome browser 107 [Internet]. 2022. Available from: http://uswest.ensembl.org/Gallus_gallus/Info/Annotation

Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, et al. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biology. 2019 Dec 30;17(1):108.

Gibbons HR, Shaginurova G, Kim LC, Chapman N, Spurlock CF, Aune TM. Divergent lncRNA GATA3-AS1 Regulates GATA3 Transcription in T-Helper 2 Cells. Front Immunol. 2018;9:2512.

Gil N, Ulitsky I. Regulation of gene expression by cis-acting long non-coding RNAs. Nat Rev Genet. 2020 Feb;21(2):102–17.

Grote P, Wittler L, Währisch S, Hendrix D, Beisaw A, Macura K, et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev Cell. 2013 Jan 28;24(2):206–14.

Guan D, Halstead MM, Islas-Trejo AD, Goszczynski DE, Ross P, Zhou H. A comprehensive prediction of transcript isoforms in 19 chicken tissues by Oxford Nanopore long-read sequencing. 2022 Jul 19;21.

Guh CY, Hsieh YH, Chu HP. Functions and properties of nuclear lncRNAs—from systematically mapping the interactomes of lncRNAs. J Biomed Sci. 2020 Mar 17;27:44.

Guo C, Gong M, Li Z. Knockdown of lncRNA MCM3AP-AS1 Attenuates Chemoresistance of Burkitt Lymphoma to Doxorubicin Treatment via Targeting the miR-15a/EIF4E Axis. Cancer Manag Res. 2020 Jul 16;12:5845–55.

Guo J, Fang W, Sun L, Lu Y, Dou L, Huang X, et al. Ultraconserved element uc.372 drives hepatic lipid accumulation by suppressing miR-195/miR4668 maturation. Nat Commun. 2018 Feb 9;9(1):612.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004 Dec;432(7018):695–716.

International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004 Dec 9;432(7018):695–716.

Jehl F, Degalez F, Bernard M, Lecerf F, Lagoutte L, Désert C, et al. RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock

Species. Frontiers in Genetics. 2021;12:1104.
Jehl F, Muret K, Bernard M, Boutin M, Lagoutte L, Désert C, et al. An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. Scientific Reports. 2020 Nov 24;10(1):20457.
Kajino T, Shimamura T, Gong S, Yanagisawa K, Ida L, Nakatochi M, et al. Divergent lncRNA MYMLR regulates MYC by eliciting DNA looping and promoter-enhancer interaction. The EMBO Journal. 2019 Sep 2;38(17):e98441.
Kern C, Wang Y, Chitwood J, Korf I, Delany M, Cheng H, et al. Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. BMC Genomics. 2018 Sep 18;19(1):684.
Kosinska-Selbi B, Mielczarek M, Szyda J. Review: Long non-coding RNA in livestock. Animal. 2020 Oct;14(10):2003–13.
Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013 Mar;14(2):144–61.
Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017 Apr 24;18(1):323.
Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet. 2017 Dec;49(12):1731–40.
Lagarrigue S, Lorthiois M, Degalez F, Gilot D, Derrien T. LncRNAs in domesticated animals: from dog to livestock species. Mamm Genome [Internet]. 2021 Nov 13; Available from: https://doi.org/10.1007/s00335-021-09928-7
Le Béguec C, Wucher V, Lagoutte L, Cadieu E, Botherel N, Hédan B, et al. Characterisation and functional predictions of canine long non-coding RNAs. Sci Rep. 2018 Sep 7;8(1):13444.
Li A, Zhang J, Zhou Z, Wang L, Liu Y, Liu Y. ALDB: a domestic-animal long noncoding RNA database. PLoS One. 2015;10(4):e0124003.
Li C, Lu L, Feng B, Zhang K, Han S, Hou D, et al. The lincRNA-ROR/miR-145 axis promotes invasion and metastasis in hepatocellular carcinoma via induction of epithelial-mesenchymal transition by targeting ZEB2. Sci Rep. 2017 Jul 5;7:4637.
Lizio M, Deviatiiarov R, Nagai H, Galan L, Arner E, Itoh M, et al. Systematic analysis of transcription start sites in avian development. PLoS Biol. 2017 Sep;15(9):e2002887.
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct;461(7265):747–53.
McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Medicine. 2014 Mar 31;6(3):26.
Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022 Apr;604(7905):310–5.
Morenos L, Chatterton Z, Ng JL, Halemba MS, Parkinson-Bates M, Mechinaud F, et al. Hypermethylation and down-regulation of DLEU2 in paediatric acute myeloid leukaemia independent of embedded tumour suppressor miR-15a/16-1. Mol Cancer. 2014 May 24;13:123.
Munschauer M, Nguyen CT, Sirokman K, Hartigan CR, Hogstrom L, Engreitz JM, et al. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. Nature. 2018 Sep;561(7721):132–6.
Muret K, Désert C, Lagoutte L, Boutin M, Gondret F, Zerjal T, et al. Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. BMC Genomics. 2019 Nov 21;20(1):882.
Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, et al. Long noncoding RNA repertoire in chicken liver and adipose tissue. Genetics Selection Evolution. 2017 Jan 10;49(1):6.
NCBI's RefSeq. Gallus_gallus-2.1 - galGal3 - Genome - Assembly - NCBI [Internet]. 2006. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_000002315.1/
NCBI's RefSeq. Gallus_gallus-4.0 - galGal4 - Genome - Assembly - NCBI [Internet]. 2011. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_000002315.3/
NCBI's RefSeq. Gallus_gallus-5.0 - Genome - Assembly - NCBI [Internet]. 2015. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_000002315.4/
NCBI's RefSeq. GRCg6a - galGal6 - Genome - Assembly - NCBI [Internet]. 2018. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_000002315.5/
NCBI's RefSeq. bGal-Gal1.mat.broiler.GRCg7b - Genome - Assembly - NCBI [Internet]. 2021a. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_016699485.2/
NCBI's RefSeq. Gallus gallus Annotation Report [Internet]. 2022. Available from: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Gallus_gallus/106/
NCBI's RefSeq N. bGal-Gal1.pat.whiteleghornlayer.GRCg7w_WZ - Genome - Assembly - NCBI [Internet]. 2021b. Available from:

https://www.ncbi.nlm.nih.gov/assembly/GCF_016700215.2
Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014 Jan;505(7485):635–40.
Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science. 2016 Jan 15;351(6270):271–5.
Pan L, Liang W, Fu M, Huang Z hua, Li X, Zhang W, et al. Exosomes-mediated transfer of long noncoding RNA ZFAS1 promotes gastric cancer progression. J Cancer Res Clin Oncol. 2017 Jun 1;143(6):991–1004.
Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare [Internet]. F1000Research; 2020. Available from: https://f1000research.com/articles/9-304
Polycarpou-Schwarz M, Groß M, Mestdagh P, Schott J, Grund SE, Hildenbrand C, et al. The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. Oncogene. 2018 Aug;37(34):4750–68.
Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009 Feb 20;136(4):629–41.
Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021 Jun 15;37(12):1639–43.
Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021 Feb;22(2):96–118.
Sun Q, Song YJ, Prasanth KV. One locus with two roles: microRNA-independent functions of microRNA-host-gene locus-encoded long noncoding RNAs. Wiley Interdiscip Rev RNA. 2021 May;12(3):e1625.
Wang H, Iacoangeli A, Popp S, Muslimov IA, Imataka H, Sonenberg N, et al. Dendritic BC1 RNA: Functional Role in Regulation of Translation Initiation. J Neurosci. 2002 Dec 1;22(23):10232–41.
Wang L, Li J, Zhou H, Zhang W, Gao J, Zheng P. A novel lncRNA Discn fine-tunes replication protein A (RPA) availability to promote genomic stability. Nat Commun. 2021a Sep 22;12(1):5572.
Wang X, Liu C, Zhang S, Yan H, Zhang L, Jiang A, et al. N6-methyladenosine modification of MALAT1 promotes metastasis via reshaping nuclear speckles. Developmental Cell. 2021b Mar 8;56(5):702-715.e8.
Wang Y, Sun L, Wang L, Liu Z, Li Q, Yao B, et al. Long non-coding RNA DSCR8 acts as a molecular sponge for miR-485-5p to activate Wnt/β-catenin signal pathway in hepatocellular carcinoma. Cell Death Dis. 2018 Aug 28;9(9):851.
Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3 Genes|Genomes|Genetics. 2017 Jan 1;7(1):109–17.
Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science. 2005 Sep 2;309(5740):1570–3.
Wright MW. A short guide to long non-coding RNA gene nomenclature. Hum Genomics. 2014 Apr 9;8(1):7.
Xu T p, Liu X x, Xia R, Yin L, Kong R, Chen W m, et al. SP1-induced upregulation of the long noncoding RNA TINCR regulates cell proliferation and apoptosis by affecting KLF2 mRNA stability in gastric cancer. Oncogene. 2015 Nov;34(45):5648–61.
Yamazaki T, Souquere S, Chujo T, Kobelke S, Chong YS, Fox AH, et al. Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. Molecular Cell. 2018 Jun 21;70(6):1038-1053.e7.
Yan C, Chen J, Chen N. Long noncoding RNA MALAT1 promotes hepatic steatosis and insulin resistance by increasing nuclear SREBP-1c protein stability. Sci Rep. 2016 Mar 3;6:22640.
Yu Y, Nangia-Makker P, Farhana L, Majumdar APN. A novel mechanism of lncRNA and miRNA interaction: CCAT2 regulates miR-145 expression by suppressing its maturation process in colon cancer cells. Mol Cancer. 2017 Sep 30;16:155.
Zhao L, Wang J, Li Y, Song T, Wu Y, Fang S, et al. NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Research. 2021 Jan 8;49(D1):D165–71.
Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015 Feb 18;16(1):97.
Zhu L, Wei Q, Qi Y, Ruan X, Wu F, Li L, et al. PTB-AS, a Novel Natural Antisense Transcript, Promotes Glioma Progression by Improving PTBP1 mRNA Stability with SND1. Mol Ther. 2019 Sep 4;27(9):1621–37.

1.3. Annoter les gènes par orthologie : forces et faiblesses de ces analyses

1.3.1. Relations évolutives entre gènes homologues et transfert de connaissances fonctionnelles

En génomique, l'orthologie fait référence à la relation évolutionnaire entre des gènes issus d'espèces différentes qui descendent d'un ancêtre commun. Des gènes orthologues sont des gènes provenant d'un même gène ancestral et ayant divergé après un événement de spéciation [225]. Par ailleurs, on distingue les gènes orthologues des gènes paralogues, ces derniers étant issus d'un phénomène de duplication au sein d'une même espèce. Dans le cas des gènes codants, les protéines produites exercent généralement des fonctions similaires dans les organismes respectifs, même si certaines fonctions peuvent être perdues ou gagnées du fait de l'évolution et que ce soit par des phénomènes de duplication des gènes (apparition de paralogues) ou de mutations [226–228]. C'est le cas par exemple du gène de la lactase qui permet la digestion du lactose chez les mammifères. Chez l'humain, une mutation est apparue permettant le maintien de l'expression de la lactase à l'âge adulte (persistance de la lactase) [229]. Cette adaptation est absente chez la majorité des mammifères, qui perdent l'expression de ce gène après le sevrage [230]. C'est également le cas pour les gènes Hox, un groupe de gènes impliqués dans le développement embryonnaire et plus précisément dans l'axe antéro-postérieur. Comme pour la plupart des gènes paralogues, la comparaison des séquences protéiques montre une plus grande similitude entre les espèces qu'au sein d'une même espèce. Rapportée aux gènes Hox, cette observation implique que ce groupe de gènes a évolué au début de l'évolution animale à partir d'un seul gène Hox par duplication en tandem et divergence ultérieure, et qu'un groupe prototypique de gènes Hox contenant au moins sept gènes Hox différents était présent dans l'ancêtre commun de tous les animaux bilatériens [231–233].

Afin de considérer le niveau de co-orthologie entre gènes de deux espèces différentes, le type de relation est usuellement [234, 235] définit comme :

- *1-to-1* (aussi noté 1:1) : signifie qu'un gène chez l'espèce de référence ne possède qu'un seul gène orthologue chez l'autre espèce ;
- *1-to-many* (aussi noté 1:m) : signifie qu'un unique gène chez l'espèce de référence possède plus d'un orthologue chez l'autre espèce dû à la présence de paralogues inexistant dans l'espèce de référence. Cela suppose un événement de spéciation, suivi d'un ou plusieurs événements de duplication chez l'autre espèce ou a fortiori chez l'un de ces ancêtres.
- *many_to_many* (aussi noté m:m) : signifie que plusieurs gènes chez l'espèce de référence possèdent plusieurs orthologues chez l'autre espèce. Ces cas sont complexes, les événements de spéciations et de duplications étant parfois compliqués à hiérarchiser.

Ainsi, malgré les possibles voies d'évolution, l'identification des relations d'orthologie reste une porte d'entrée intéressante en génomique comparative, car elle permet de transférer des connaissances fonctionnelles entre espèces par extrapolation et permet soit de s'affranchir, soit d'orienter une partie des démarches expérimentales afin de compléter l'annotation des génomes. Plus globalement, elle aide également à reconstruire l'histoire évolutive des gènes et des génomes.

1.3.2. Prédiction de l'orthologie des gènes codant des protéines par des approches bio-informatiques intégratives

Il existe plusieurs méthodes bio-informatiques pour prédire l'orthologie, notamment pour les PCG. Les méthodes basées sur la similarité de séquences comparent toutes les paires de protéines entre les génomes pour détecter les meilleures réciprocités, correspondant généralement à des orthologues. Malgré l'existence de plusieurs algorithmes d'alignement tels que BLAT [236], UBLAST [237], LAST [238] ou encore DIAMOND [239] et MMseqs [240], BLAST (*Basic Local Alignement Search Tool*) [94, 241] et ses dérivés (spécifiquement dans ce cas blastp) restent probablement l'algorithme heuristique local le plus utilisé dans la quête des orthologues [242], particulièrement lors d'un usage en indépendant. Cependant, dans le cadre d'intégration dans les *pipelines*, DIAMOND semble être utilisé de manière plus récurrente. C'est par exemple le cas pour les logiciels comme InParanoid [243], OMA [244] ou

encore OrthoFinder [245] qui utilisent en complément des algorithmes d'alignement des approches par *clustering* pour construire des familles de protéines orthologues.

D'autres approches reposent sur la synténie, c'est-à-dire la conservation de l'ordre des gènes entre espèces. Elle constitue un signal fort pour déduire/appuyer l'orthologie puisque les gènes qui descendent du même gène sont susceptibles de faire partie d'un bloc de gènes, tous dans le même ordre, dans les deux espèces. Certains réarrangements entre les gènes peuvent se produire au fil du temps, mais la conservation de l'environnement reste un point à considérer [246]. Des méthodes comme OrthoCluster [247, 248] analysent la synténie pour prédire les relations d'orthologie. Ensembl-Compara utilise par exemple la synténie pour produire un score de conservation de l'ordre des gènes (« *Gene Order Conservation score* ») afin de déterminer la probabilité que les paires d'orthologues identifiées soient belles et bien confirmées [249].

Les arbres phylogénétiques peuvent également venir en complément. La topologie d'un arbre permet de distinguer spéciation et duplication, les vrais orthologues étant séparés par un nœud de spéciation. Des algorithmes comme TreeFam [250] ou LOFT [251] utilisent les informations phylogénétiques pour démêler orthologie et paralogie. De ce fait, la plupart des approches combinent différents critères pour une prédiction plus fiable. OrthoFinder [245] ou eggNOG [252] intègrent à la fois similarité de séquences, synténie et phylogénie pour inférer l'orthologie. L'intégration de données multiples pallie les limites de chaque méthode prise ainsi isolément. Avec l'accumulation des données génomiques, de puissantes bases de données d'orthologues couvrant de larges portions du vivant ont vu le jour. OMA [244, 253, 254] et eggNog [252, 255] rassemblent aujourd'hui plus de 5 000 génomes eucaryotes et procaryotes, OrthoDB [256, 257] permet de naviguer parmi 100 millions de gènes prédits et leurs potentiels liens d'orthologie dans près de 2 000 génomes eucaryotes, 18 000 procaryotes et 8 000 virus. La base de données de référence Ensembl propose, elle aussi, des listes de PCG orthologues, notamment par son API Compara [258] ou par l'usage de BioMart [259]. Ces ressources facilement consultables grâce aux interfaces Web et API proposées constituent des références pour transférer les connaissances entre espèces et s'intéresser à l'évolution des gènes.

Ainsi, les études d'orthologie à grande échelle, portant sur des centaines d'espèces, ont permis de nombreuses avancées en biologie évolutive. Seule une petite fraction des gènes serait apparue plus récemment, après la divergence des grands groupes taxonomiques. Par

exemple, on estime que 70 % des gènes PCG humains ont des orthologues clairement identifiables chez des espèces très éloignées comme le poisson-zèbre, soulignant la forte conservation du génome des vertébrés sur plus de 450 millions d'années d'évolution [260].
En conclusion, l'inférence d'orthologie à haut débit est devenue possible grâce à des méthodes computationnelles combinant diverses sources de données génomiques. Les relations prédites doivent cependant être considérées avec précaution, un examen manuel restant nécessaire pour les analyses fonctionnelles critiques. L'intégration de nouveaux types de données comme l'expression des gènes ou les interactions protéines-protéines devrait améliorer encore la fiabilité des prédictions.

1.3.3. Déterminer l'orthologie des gènes à l'ère des lncRNA : complexités et obstacles

Au cours du temps et face aux différentes contraintes, les gènes évoluent, que ce soit en structure ou en fonction. Ces changements, dont l'impact peut varier, mettent en jeu de nombreux mécanismes moléculaires et évolutifs qui peuvent rendre les détections par orthologie complexes. Au niveau de leur séquence, les gènes sont composés de différentes structures aux caractéristiques propres, notamment les exons qui codent en partie pour les protéines (hors UTR – *untranslated regions*), les introns non-codants, mais également les régions régulatrices en amont et en aval comme les UTR, les promoteurs proximaux ou encore les *enhancer*/*silencer*. Chacune de ces parties est soumise à des pressions de sélection particulières au cours de l'évolution [261].
Plus précisément, la partie des exons codant pour les protéines est soumise à une pression de sélection purificatrice forte, qui élimine la majorité des mutations délétères changeant la séquence en acides aminés de la protéine. Cependant, certaines mutations ponctuelles peuvent être sélectionnées positivement si elles confèrent un avantage sélectif [262]. Des changements plus importants comme des insertions ou des délétions d'exons entiers peuvent aussi survenir, modifiant plus radicalement la protéine. Ce mécanisme peut permettre l'émergence de nouvelles fonctions par néo-fonctionnalisation. Les introns sont a priori dépourvus de fonction codante, ils évoluent donc plus rapidement que les exons, car soumis à une pression de sélection relâchée. Cependant, les extrémités des introns jouent un rôle crucial lors de l'épissage alternatif, permettant la production de multiples isoformes

protéiques à partir d'un même gène et peuvent avoir des rôles de régulation de l'expression des gènes, leur conférant une valeur adaptative [263]. Les mutations modifiant les sites d'épissage sur les introns peuvent donc avoir des effets phénotypiques importants [264]. De plus, l'épissage alternatif est finement régulé selon le type cellulaire ou le stade de développement, conférant une expression spatio-temporelle précise aux différentes isoformes [265]. Les régions non traduites en 5' et 3' des gènes (UTR) contiennent également des séquences régulatrices clés. Les UTR en 5', ainsi que les promoteurs proximaux en amont de ces derniers, peuvent comporter des motifs fixant les facteurs de transcription tels que le motif de « TATA box » et autres, qui déterminent quand et où le gène sera exprimé [266, 267]. Les mutations de ces séquences *cis*-régulatrices changent les profils d'expression des gènes, fournissant une source majeure de variabilité phénotypique [268]. Les UTR en 3', qui sont en général beaucoup plus longs que les 5'UTR, contiennent des séquences modulant la stabilité des ARNm ou leur traduction, qui peuvent par ailleurs évoluer [269]. D'autres mécanismes évolutifs agissent à plus grande échelle, venant altérer l'organisation des gènes au sein des génomes. Les duplications, délétions ou réarrangements de segments génomiques entiers sont des moteurs puissants d'innovation évolutive [270]. En effet, les réarrangements changent l'organisation chromosomique par inversion ou translocation, pouvant modifier la régulation des gènes [271]. L'évolution des éléments transposables est également un moteur clé de l'évolution des génomes et des gènes. Les transposons sont des séquences d'ADN mobiles qui peuvent s'insérer dans les gènes, modifiant leur séquence ou leur expression [86]. Bien que souvent délétères, ces insertions fournissent parfois de nouveaux sites d'épissage, promoteurs ou *enhancers* modulant l'activité des gènes de façon adaptative [272, 273]. L'évolution des gènes est donc un processus multi-échelle soumis à diverses pressions de sélection et intégrant divers mécanismes, que ce soit les mutations ponctuelles dans les séquences codantes et régulatrices, l'évolution des profils d'épissage alternatif, les réarrangements chromosomiques ou encore l'activité des éléments transposables. Impliquant plus ou moins une combinaison de ces processus, au cours de l'évolution, certains gènes codants peuvent ainsi se transformer en pseudogènes puis en ARN longs non-codants (lncRNA) [274]. Les pseudogènes dérivent de gènes codants qui ont perdu leur capacité à produire une protéine fonctionnelle, le plus souvent par accumulation de mutations non-sens ou décalant le cadre de lecture. Même si cela reste à l'étude, certains pseudogènes (*e.g.*, PTENP1 agissant conjointement avec le PCG PTEN) peuvent cependant acquérir de nouvelles

fonctions régulatrices : leurs transcrits agissent alors en tant que leurres moléculaires séquestrant des microARN ou des protéines, modulant l'expression d'autres gènes [275] ou via d'autres mécanismes décrits en §1.2.3.1 pour les lncRNA. En effet, des études évolutives suggèrent que de nombreux lncRNA dérivent de gènes codants qui ont été cooptés pour de nouvelles fonctions régulatrices au cours de l'évolution [276]. L'apparition de mutations prématurées de terminaison de la traduction ou de sites de polyadénylation alternatifs peut générer des transcrits tronqués qui échappent à la traduction sans affecter les fonctions régulatrices naissantes. Plusieurs mécanismes génèrent donc une dichotomie évolutive entre la séquence codante d'un gène ancestral et ses fonctions régulatrices potentielles. Cette séparation permet l'émergence de nouveaux ARN non-codants à partir de gènes codants, fournissant une source d'innovation évolutive [277]. L'inactivation progressive de la séquence codante aboutit à la formation de pseudogènes, tandis que le maintien de fonctions régulatrices donne naissance à des lncRNA. À titre d'exemple, le pseudogène PTENP1 dérive du gène suppresseur de tumeur PTEN mais a acquis des fonctions de régulation post-transcriptionnelle de PTEN et d'autres gènes par séquestration de microARN [278]. Le lncRNA XIST, essentiel à l'inactivation du chromosome X chez les mammifères, semble dériver d'un gène codant ancestral, LNX3 [279, 280] dont il a conservé des séquences régulatrices en *cis* [281]. Le lncRNA HOTAIR, impliqué dans la régulation épigénétique, provient de la cooptation d'un gène codant HOXC en lncRNA régulateur [282]. Finalement, les lncRNA présentent généralement une séquence primaire peu conservée entre espèces, contrairement aux gènes codants soumis à une forte pression de sélection au niveau protéique [276, 283]. Cette divergence rapide rend délicate l'identification de lncRNA orthologues sur la base de similarité de séquence. De plus, l'expression des lncRNA apparaît très spécifique du tissu et du stade développemental, ce qui complique la comparaison de leur expression entre organismes [284]. De plus, les lncRNA sont fréquemment soumis à des réarrangements génomiques qui brouillent leur histoire évolutive. Translocations, inversions, insertions et délétions rendent difficile l'identification de synténie conservée, un critère clé pour définir l'orthologie [283]. Tout comme les PCG et de par leur nature régulatrice, les lncRNA sont également sensibles aux modifications des séquences *cis*-régulatrices adjacentes, dont la position n'est pas nécessairement conservée au cours de l'évolution [285, 286]. Par ailleurs, les mécanismes de duplication des lncRNA sont mal connus. Alors que la duplication segmentale ou en tandem est bien décrite pour les gènes codants, on ignore dans quelle

mesure elle s'applique aux lncRNA [287]. Cette duplication extensive brouille l'identification d'orthologues 1:1 entre espèces. S'y ajoute la présence de nombreux éléments transposables au sein des lncRNA, source fréquente de duplication et de réarrangements locaux [288]. Enfin, le manque d'annotations exhaustives et standardisées des lncRNA dans de nombreux génomes limite les comparaisons globales [114]. Malgré ces défis, plusieurs approches bio-informatiques émergent pour démêler l'évolution des lncRNA [103]. L'analyse phylogénétique de séquences conservées au sein de lncRNA permet d'identifier des domaines fonctionnels soumis à sélection. La recherche de motifs structuraux conservés, de sites de fixation à des protéines ou microARN, ou de promoteurs partagés, fournit des indicateurs de conservation fonctionnelle [276, 283]. Enfin, la comparaison des profils d'interaction et de régulation de l'expression génique entre orthologues peut prédire des fonctions équivalentes. Associées à des validations expérimentales, ces approches devraient aider à esquisser peu à peu le répertoire des lncRNA orthologues entre espèces.

Ainsi, au vu des mécanismes d’évolutions des gènes et de leurs conséquences, établir des relations d'orthologie pour les lncRNA peut donc s'avérer difficile comparé aux gènes codants. Bien que complexe, l'analyse évolutive des lncRNA apporte cependant un éclairage crucial sur l'émergence de fonctions régulatrices innovantes au cours de l'évolution des génomes et des transcriptomes et participe à une meilleure annotation de ces derniers.

2. Mesurer l'expression des gènes par séquençage ARN

2.1. Le séquençage ARN et ses différents usages

2.1.1. Principes, méthodes et analyses

Le séquençage ARN (abrégé en RNAseq pour *RNA sequencing*) est une technique permettant d'analyser et de quantifier l'ensemble des ARN transcrits dans un échantillon biologique (tissu ou cellule), appelé transcriptome [289]. Cette approche, considérée maintenant comme référence pour quantifier un transcriptome, à évoluer tant en termes de coûts que de techniques depuis son apparition il y a une quinzaine d'année [104, 290, 291]. Contrairement aux puces à ADN qui ne permettent d'analyser que certains gènes définis en amont, le RNAseq est sans a priori et permet, dépendamment de la profondeur, de détecter tous les transcrits présents dans l'échantillon d'intérêt, y compris les ARN non codants [103, 292] ou encore les ARN circulaires [293].

Plusieurs étapes sont nécessaires pour réaliser le RNAseq. Tout d'abord, les ARN totaux de l'échantillon biologique d'intérêt, que ce soit des cellules en culture, des tissus ou des organismes entiers, doivent être extraits. Par la suite, il est possible de sélectionner certains types de transcrits selon les analyses souhaitées. En effet, la majorité des ARN d'un échantillon appartiennent à la classe des ARN ribosomiques (rRNA) ou de transferts (tRNA) mais il est possible de s'en affranchir en sélectionnant par exemple les ARN présentant une queue poly-A (signe des ARN matures) ou encore les ARN après ribo-déplétion (élimination des rRNA). Il est également possible de sélectionner les ARN selon leur taille, ce qui est couramment utilisé pour l'analyse des micro-ARN comparativement aux autres ARN [294, 295]. Il est critique d'utiliser des protocoles optimisés pour obtenir des ARN de bonne qualité et éviter des biais dans les résultats. Les ARN extraits sont ensuite convertis en ADNc (ADN complémentaires) grâce à une transcriptase, puis fragmentés de manière aléatoire en fragments de 300 à 500 pb dans le cadre du séquençage *short-read*. Des adaptateurs sont ajoutés aux extrémités des fragments d'ADNc qui sont ensuite amplifiés par PCR et séquencés massivement grâce aux technologies de séquençage haut-débit [295].

Deux stratégies majeures de séquençage co-existent actuellement : *i)* La plus répandue et la moins coûteuse, le séquençage *short-read* qui génère des courtes lectures de 50 à 250 pb en moyenne correspondant aux extrémités des fragments d'ADNc préalablement amplifiés, et *ii)* la plus récente, le séquençage *long-read* produisant des lectures continues de quelques

kilobases (kb) à plusieurs dizaines de kb. Le *short-read*, principalement réalisé grâce à des appareils de type Illumina, leader sur le marché avec plus de 80% des parts dans le domaine du séquençage, présente l'avantage d'avoir un débit extrêmement élevé, générant des centaines de millions de lectures pour un coût faible compris entre 150 et 350€ environ [296, 297]. Cependant, l'assemblage bio-informatique des courtes lectures est complexe. Le *long-read,* plus récent et proposé par des entreprises telles que Pacific Biosciences (PacBio) [298] ou Oxford Nanopore Technologies (ONT) [299], permet de séquencer des transcrits entiers. Il facilite l'assemblage, notamment au niveau des régions répétées ou à faible complexité, mais a un débit plus faible et un taux d'erreur plus élevé [300]. Les deux approches apparaissent donc complémentaires pour les multiples usages.

Une fois le séquençage effectué, les données biologiques brutes prennent généralement la forme de données numériques sous la forme de fichiers *.fastq* contenant les séquences nucléotidiques et les scores de qualité pour chaque lecture [301]. L'analyse bio-informatique comporte par la suite plusieurs étapes. Tout d'abord, un contrôle qualité permet d'éliminer les séquences de mauvaises qualités ou contaminantes. Les *reads* sont ensuite alignées sur le génome de référence de l'espèce en question. Si l'espèce est mal annotée ou ne possède pas de génome de référence, l'assemblage dit « *de novo* » du transcriptome est réalisée [302–305]. Le but est de savoir l'origine du *read* en termes de positions dans le génome et en prenant en compte l'orientation du *read* originel, le transcrit étant mono-brin. Des outils spécifiques du RNAseq comme STAR [306, 307] (le plus utilisé), HISAT [308, 309] ou TopHat (le plus ancien) [310, 311] sont couramment utilisés pour aligner les données *short-read*, tandis que Minimap2 [312–314] convient mieux pour les *long-reads* grâce à une vitesse d'exécution et une précision plus importante que d'autres algorithmes [315] tels que NGMLR [316, 317], GraphMap [318, 319] ou encore LAMSA [320, 321]. Notons que les outils couramment utilisés pour les données issues du séquençage ADN (abrégé en DNAseq pour *DNA sequencing*) ne peuvent pas être utilisés tels quels pour les données RNAseq. En effet, les *reads* de RNAseq sont issus de transcrits matures et donc épissés et peuvent en conséquence s'aligner sur deux ou plusieurs exons qui sont, à l'échelle du génome, séparé par une partie intronique pouvant atteindre plusieurs kilobases. Les outils spécifiques du RNAseq doivent en conséquence laisser la possibilité de « découper » les *reads* de manière à intégrer les informations exoniques et introniques pour ensuite les aligner de façon optimale.

2.1.2. Les différentes applications du RNAseq

Dans un premier temps, le RNAseq permet de modéliser de nouveaux gènes, mais également de préciser les extrémités UTR des PCG généralement modélisés dans les années 2000 comme pour l'humain ou la poule. Il permet de plus de préciser la structure des gènes en détectant de nouveaux exons et introns issus d'épissages alternatifs. Grâce à l'émergence du RNAseq, la connaissance des loci géniques a donc été grandement améliorée comme le démontre le contenu des différentes versions des bases de données de référence telles que Ensembl et NCBI ou encore des projets plus spécifiques tels que FR-AgENCODE [322, 323], projet intégré au sein du consortium FAANG [195, 196] cherchant à améliorer les annotations des génomes des espèces d'élevages. C'est grâce à ces données de séquençage RNA que de nombreux pseudogènes et lncRNA ont ainsi pu être mis en évidence dans des régions auparavant considérées comme non transcrites. Cependant, les séquences utilisées à ces fins étant principalement du *short read*, la connaissance des transcrits associés aux différents loci géniques s'avèrent encore imparfaite.

Une autre finalité du RNAseq est l'obtention des niveaux d'expression des gènes. Le RNAseq permet ainsi de caractériser et de comparer de façon quantitative les transcriptomes de différents échantillons biologiques par l'intermédiaire de l'identification des gènes différentiellement exprimés (noté DEG pour « *Differentially expressed genes* ») entre des conditions [124, 125, 324, 325], les logiciels les plus populaires étant DESeq2 [326, 327] ou edgeR [328, 329]. Associé ou non à d'autres données omiques, le RNAseq aide également à reconstruire les réseaux de régulation de l'expression des gènes et à comprendre comment l'expression génique est régulée [292, 330]. Pour ce faire, le logiciel WGCNA [331] est le plus usuellement utilisé. Par ailleurs, le RNAseq fournit une mesure de l'expression des gènes plus précise que les puces à ADN qui étaient anciennement utilisées [332]. Ainsi, le RNAseq est utilisé à des fins de quantification du transcriptome de tissus dans de nombreux domaines de la biologie. Par exemple, dans le domaine de la santé, le RNAseq sert à identifier des biomarqueurs d'états pathologiques comme les cancers, pour découvrir de nouvelles cibles thérapeutiques, ou encore pour étudier les mécanismes moléculaires de maladies génétiques. En recherchant les transcrits anormalement exprimés dans des tissus cancéreux comparativement à des tissus sains, des signatures d'expression caractéristiques de cancers ou de leurs stades ont pu être établies [333, 334]. Le RNAseq est également très utile en

immunologie et microbiologie pour caractériser finement les réponses des cellules immunitaires ou les transcriptomes de micro-organismes dans diverses conditions environnementales ou d'infection [335, 336]. Enfin, de par son étape de séquençage, le RNAseq permet d'identifier des variants génétiques présents dans l'ADN génomique, mais aussi dans les ARN indépendamment du génome, processus appelé édition des ARN (*editing*) [337–341]. Ces variations spécifiques de l'ARN sont un des mécanismes de maturation des transcrits qui peut parfois conduire à différentes maladies [342, 343]. Cependant, l'édition des ARN est un processus assez rare, peu d'événements d'édition sont généralement identifiés que ce soit chez la souris [344], la poule (<200 événements) [345–347] ou l'humain (<1000 événements) [348, 349]. De plus, ce phénomène est plus fréquent dans les régions répétées et donc peu représenté dans les *reads* alignés [350]. Ainsi, en considérant l'édition comme un phénomène marginal, le RNAseq permet également d'identifier des variants génétiques présents dans l'ADN génomique [337]. Cette application est cependant moins populaire que l'utilisation à des fins de quantification du transcriptome. Pour finir, le RNAseq permet l'étude de l'expression allèle-spécifique (ASE, pour *Allele Specific Expression*), qui nécessite à la fois l'étude de l'expression et la détection de variant hétérozygotes [351, 352].

En résumé, grâce à sa capacité à séquencer les parties exprimées du génome que sont les transcrits, le RNAseq est devenu très populaire dans la communauté scientifique pour explorer les mécanismes fins de la régulation de l'expression des gènes, les impacts de traitements ou conditions ou encore pour comprendre les relations entre génotype et phénotype comme nous le développerons en §3.2.

2.2. Quantification de l'expression génique et prise en compte de la diversité des contextes

2.2.1. Différentes métriques de quantification de l'expression génique par RNAseq : forces et faiblesses

En préambule, notons que la quantification de l'expression des gènes est dépendante de l'annotation de référence utilisée pour le génome de l'espèce considérée, ainsi seuls les modèles de gènes identifiés sur le génome sont généralement quantifiés.

Dans le cadre des données issues de *short-read*, la métrique élémentaire est le « *raw counts* » (ou *total counts*) qui permet le décompte brut des *reads* alignés sur un modèle génique. Cependant, la distribution non normale de cette métrique, caractérisée par beaucoup de valeurs faibles, pose un problème pour de nombreux tests statistiques. De plus, de par une interprétation biologique compliquée et du fait de la variabilité entre échantillons du nombre de *reads* séquencés (biais de librairies) cette métrique élémentaire ne peut donc pas être utilisée pour les analyses ultérieures visant à comparer différents échantillons. Notons qu'il y a également un biais dû à la variation de la taille des gènes, les gènes longs captant plus de *reads* que les courts gènes. Bien que ce biais n'impacte pas les comparaisons entre échantillons, certaines métriques sont proposées pour prendre en compte à la fois le biais de librairies, mais aussi le biais de taille des gènes. Pour corriger le biais de librairies, l'hypothèse faite est que la majorité des gènes soient également exprimés entre échantillons ; une somme de *reads* équivalente entre ces derniers est donc attendue.

Ainsi, parmi les métriques normalisées les plus courantes, on retrouve *les reads per kilobase per million mapped read*s (RPKM) et/ou *les fragments per kilobase per million mapped reads* (FPKM) [290]. Ces deux métriques analogues (le FPKM correspond au RPKM pour des données de RNAseq dites *paired-end*) normalisent le nombre de *reads* alignés sur un gène par rapport à sa longueur et à la profondeur de séquençage de la librairie, permettant ainsi de comparer l'expression de gènes de tailles différentes entre eux et entre échantillons. Sa formule est donc la suivante :

$$RPKM_G = \frac{R_G}{R_L \times L_G}$$

$RPKM_G$ = valeur de RPKM pour le gène G ; R_G = nombre de reads alignés pour le gène G (en millions – 10^6) ; R_L = nombre de reads totaux alignés pour la librairie L (en millions – 10^6) ; L_G = Longueur (en kbp – 10^3 bp) du gène G.

Simples à calculer, les RPKM/FPKM ont été les premières métriques normalisant les données brutes *raw counts* et ont largement contribué à la démocratisation du RNAseq. Néanmoins, ces métriques possèdent quelques faiblesses pouvant apporter des résultats inconsistants, notamment pour l'analyse de gènes différentiellement exprimés. L'une d'elle est de ne pas prendre en compte le fait qu'une faible proportion de gènes représente une grande proportion de *reads* et va donc fortement impacter, à tort, les facteurs de normalisation pour le biais de librairie [353]. Par conséquent, une autre métrique de données normalisées, est apparue pour pallier ce dernier point : *le transcript per million* (TPM) [354, 355]. Comme le RPKM, le TPM normalise les *raw counts* par la taille du gène, mais prend en compte les variations de librairies via les variations du nombre total de *reads* ramenés à l'échelle du transcrit. Par construction, la somme des TPM est constante et égale à 1 million pour tous les échantillons. Il est calculé comme suit :

$$TPM_G = \frac{R_G/L_G}{\sum_G R_G/L_G} = \frac{RPKM_G}{\sum_G RPKM_G}$$

TPM_G = valeur de TPM pour le gène G ; R_G = nombre de reads alignés pour le gène G (en millions – 10^6) ; L_G = Longueur (en kbp – 10^3 bp) du gène G

De par sa propriété d'obtention d'une somme constante entre échantillons, le TPM a pris le pas sur les RPKM et de nombreux logiciels de quantification d'expression tels que RSEM [355], Kallisto [356] ou Salmon [357] proposent en sortie les données brutes en *counts* et les données normalisées sous forme de TPM.

Les métriques précédentes ne prennent pas en compte le fait que qu'une faible proportion de gènes représente une grande proportion de *reads.* Notons par exemple que 25 % des PCG les plus exprimés correspondent à plus 80 % des *reads* alignées (voir *review-1*) [103]. Ces gènes, s'ils sont variables biologiquement, vont donc fortement impacter, à tort, la normalisation pour le biais de librairie. Ainsi, de nouvelles métriques prenant en compte ce problème ont vu le jour. Tout d'abord, le « *Trimmed Mean of M-values* » (TMM), qui élimine 5 % de gènes les plus exprimés avant de calculer le facteur de normalisation. Par ailleurs, comme cette métrique est très utilisée en analyse différentielle (étude de N conditions), elle retire également les gènes les plus différentiellement exprimés (pour TMM, 30 % des top gènes les plus différentiellement exprimés) car ils impacteraient à tort la normalisation. La méthode TMM va calculer un facteur de normalisation pour chaque échantillon par comparaison aux

autres échantillons sur les gènes filtrés comme indiqués plus haut, ce facteur devant être en théorie de 1 s'il n'y a pas de biais. La démarche suivie est plus précisément exposée dans l'article de Robinson et al., 2010 [358]. Cette méthode de normalisation est par ailleurs implémentée dans le package edgeR Bioconductor [329] comme méthode de normalisation par défaut.

Proche du TMM, la méthode de normalisation du « *Relative Log Expression* » (RLE) est également d'usage et est proposée dans le package DESeq2 Bioconductor [326, 359]. Basée sur les mêmes principes que la méthode TMM, elle calcule un facteur de normalisation par échantillon, calcul fait sur un set de gènes qui n'est ni le « top » gènes exprimés ni le « top » gènes DE entre conditions [360].

Selon le dispositif expérimental, il est également possible d'ajuster les modèles en prenant en compte des artefacts techniques connus ou inconnus dans le modèle statistique approprié en fonction de la question posée. Il est de plus d'usage d'observer les données par ACP (pour *Principal Component Analysis*) [361] pour détecter d'éventuels effets connus. Une explication illustrée (mais non exhaustive) des différents processus pouvant être appliqués pour la normalisation est proposée par le créateur de contenu Josh Starmer dans une série de vidéos dédiés [362–365].

Comme évoqué précédemment, l'utilisation de données *long-read* se répand, mais certains challenges se présentent également, notamment pour estimer les niveaux d'expression des gènes [366] en raison d'un nombre de reads généralement beaucoup plus faible [367, 368] ainsi qu'un taux d'erreur de séquençage élevé (10-15 %). Ainsi, des outils spécifiques ont été développés pour permettre d'estimer l'abondance des transcrits à partir de ce type données de séquençage [369–371], cependant cela reste encore un champ de recherche où les outils ne sont pas encore stabilisés. Notons que peu de données de ce type sont à ce jour disponibles chez la poule [372].

En résumé, le RNAseq a conduit au développement de différentes métriques permettant de normaliser l'expression génique, chacune avec leurs forces et faiblesses. Le choix de la métrique dépend donc de la question biologique avec, par exemple, un usage préférentiel du TMM et RLE pour des analyses DE alors que le TPM s'avère davantage populaire pour l'analyse de données issues de tissus et de projets variés. Par ailleurs, de nombreux états de l'art proposent de manière régulière une vue d'ensemble de ces métriques pour les utiliser dans les meilleures conditions [360, 373–376].

2.2.2. Mesures de la tissu-spécificité

Les données d'expression géniques issues du RNAseq sont très souvent utilisées pour comparer les niveaux d'expression des gènes entre différents contextes biologiques et pour mieux comprendre les fonctions biologiques impactées par ces contextes. Il s'agit, par exemple, de comparer différents stades développementaux ou encore différentes conditions expérimentales. Ces études sont portées dans un ou des tissus ou types cellulaires dans lesquels est supposé un impact des conditions testées sur les transcrits. Connaitre la spécificité (ou non) d'expression d'un gène, c'est-à-dire son expression préférentielle (ou non) dans un ou des type(s) cellulaire(s), est donc une information clef pour notamment orienter les hypothèses quant au tissu à étudier. Ainsi, afin d'évaluer la spécificité tissulaire des gènes, il est nécessaire d'avoir un jeu de données multi-tissues cependant si plusieurs indicateurs existent, il n'existe pas de méthodes de référence [377].

Les métriques de tissus-spécificité

Il est possible de distinguer trois types de métriques. D'une part, les métriques telles que le tau de Yanai [378], le TSI (*tissue specificity indices*) [379] ou encore le Hg de Schug et al. [380], qui donnent un unique indicateur par gène pour l'ensemble des tissus. D'autre part, des métriques telles que le PEM (*Preferential Expression Measure*) [381] ou encore le z-score [382] qui produisent un indicateur pour chaque tissu permettant d'avoir une vue spécifique pour chacun d'entre eux. Notons que le z-score est notamment employé dans des domaines autres que la tissue spécificité. Ces métriques sont en réalité déjà très dépendantes du type de normalisation appliquée pour l'expression des gènes et présentent chacune leurs forces et faiblesses telles que détaillés par Kryuchkova-Mostacci et Robinson-Rechavi en 2017 [377]. Enfin, une autre métrique, la plus simple, est le *fold-change*, qui mesure le rapport d'expression d'un gène entre le premier et le second tissu dans lesquels le gène est le plus exprimé, un gène sera alors considéré comme tissu-spécifique si ce *fold-change* est supérieur à un certain seuil.

Les seuils utilisés

Pour ce qui est de la métrique tau qui est la plus usuellement employée, elle fournit un score global allant de 0 pour les gènes parfaitement ubiquitaires à 1 pour les gènes strictement

spécifiques. Brièvement, cette métrique considère l'expression d'un gène dans chaque tissu et calcule un ratio par rapport à l'expression maximale. Une moyenne de ces ratios est alors calculée et donne un indicateur de spécificité.

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1} \; ; \; \hat{x}_i = \frac{x_i}{\max\limits_{1\leq i\leq n} x_i}$$

x_i = valeur d'expression dans le tissue i ; n = nombre de tissus

Si encore une fois aucun seuil n'est clairement défini, un gène présentant un tau supérieur à 0.8 [383, 384] est souvent considéré comme tissu-spécifique (noté TS) même si la valeur de 0.9 semble de plus en plus utilisée [385, 386]. Par ailleurs, si cette métrique est sensible au choix d'appliquer ou non une transformation logarithmique à l'expression des gènes en amont, elle est également par construction sensible au nombre de tissus étudiés et à la diversité de ces tissus [124, 125]. En effet, si l'on considère un dispositif multi-tissues constitué d'un sous-groupe de tissus ayant des caractéristiques proches (*e.g.,* des sous-tissus du cerveau), un gène peut ne pas être TS, contrairement à ce qui est attendu. En effet, le poids de ce sous-groupe dans le dispositif peut être tel qu'il impacte trop fortement la moyenne inter-tissu. A contrario, un gène peut être TS, mais définir le nombre de tissus qui se détachent du tissu moyen nécessite alors d'appliquer un seuil arbitraire. Si la première approche consiste ainsi à ne considérer que le tissu le plus exprimé, Lüleci et al. ont proposé une extension de la métrique tau (nommée *tau-extended)* afin d'affiner cette sélection [384]. Brièvement, un intervalle d'expression par rapport à l'expression maximale et considérant la variabilité d'expression des tissus et donc définie et chaque tissu présentant une expression incluse dans cet intervalle est ensuite considéré comme spécifique.

3. Approcher les caractères complexes par l'étude de l'expression génique

3.1. Variants et variations des caractères complexes

3.1.1. Le modèle polygénique additif infinitésimal

Les caractères ou maladies d'intérêt sont souvent des caractères dits complexes ou quantitatifs. Ils ont généralement une variation continue qui suit une loi gaussienne. Ces caractères sont impactés par la génétique (G) et l'environnement (E). La composante génétique G est définie selon le modèle polygénique infinitésimal publié par Ronald Fisher en 1918 et en lien avec le théorème central limite [387]. Ainsi, il est considéré que le phénotype (P) d'un individu est le résultat de son génotype (G) et de son environnement contrôlé ou non (E). La valeur génétique G (dite additive) se définit comme la somme des petits effets des deux allèles d'une infinité de loci gouvernant le caractère d'intérêt et composant l'individu. Notons que les valeurs génétiques de dominance et d'épistasie sont en général négligées faute de ne pouvoir les estimer proprement à l'échelle du génome. De même, l'interaction des deux composantes génétiques et d'environnement (GxE) est également négligée. On peut ainsi écrire, à l'échelle de l'animal :

$$P = G + E + [G \times E]$$

De même, à l'échelle de la population :

$$V(P) = V(G) + V(E) + [Cov(G \times E)]$$

La variation $V(x)$ correspond à la variance de x et la variance d'un effet combiné $Cov(x)$ à la covariance. Pour un caractère d'intérêt, il est intéressant d'estimer l'héritabilité h^2, qui mesure la part de la variabilité génétique impliquée dans la variabilité phénotypique observée, et est donc définie comme suit :

$$h^2 = \frac{V(G)}{V(P)}$$

Ce ratio, compris par construction entre 0 et 1, est relatif à un caractère précis dans une population donnée. Afin d'avoir quelques ordres de grandeurs, notons qu'un caractère est

considéré comme peu héritable pour un $h^2 < 0,2$ (*e.g.,* caractères liés à la reproduction) et fortement héritable pour un $h^2 > 0,4$ (*e.g.,* composition corporelle). Cette héritabilité et plus précisément sa composante génétique V(G) relève donc à l'échelle moléculaire de l'effet d'une infinité de variations génétiques dans le génome avec des petits effets sur le caractère, mais variables entre allèles. Il convient alors d'évaluer le nombre de variations dans le génome, ce qui devenu possible depuis quelques années grâce aux technologies de séquençage à haut débit, mais également de localiser celles impliquées dans la variation des caractères d'intérêt. Pour finir, l'ultime étape, encore bien laborieuse, consiste à connaitre leur mécanisme d'action.

3.1.2. Une infinité de variants dont une poignée responsable de la variation des caractères complexes

Grâce au développement des technologies de séquençage haut débit, il est devenu possible de mieux caractériser les variations du génome d'une espèce donnée. Une étude de 2015 menée par le « 1000 Genomes Project Consortium » estime qu'un génome humain, d'une taille avoisinant 3 Gb, comporte environ 4 à 5 millions de polymorphismes en moyenne par individu [388]. Dans cette étude, incluant plus de 2 500 individus et représentant 26 populations, un total de plus de 88 millions de variants ont été détectés. Cependant, plus des 3/4 possèdent une fréquence à l'échelle de la population (MAF pour *Minor Allele Frequency*) inférieure à 0,5 % soulignant la nature rare de ces derniers [388]. Plus de 90 % de ces variants sont en réalité des substitutions d'un seul nucléotide appelées SNP (pour *Single Nucleotide Polymorphism*), les 10 % restant étant associés à de courtes insertions/délétion (INDEL pour *INsertion DELetion*). De manière rare, il a été également observé des variants structuraux affectant de larges segments chromosomiques. De même, l'étude portée par le consortium GTEx [389] indique l'identification de 66M de SNP dont 18M avec une MAF ≥ 1 % au travers de 838 individus. La base de données de SNP, dbSNP d'Ensembl, reporte à ce jour (V110 – juillet 2023) plus de 700M de SNP/INDEL chez l'humain [390]. Concernant la poule dont la taille du génome atteint environ 1 Gb, un total de 24,5M de SNP et de 1,3M d'INDEL est reporté [391]. Quant à la base de donnée Galbase [392] incluant 928 individus issus de 47 races domestiques et cinq sous-espèces de la race ancestrale *Red Jungle Fowl*, elle dénombre plus de 21,5M de SNP et 2,7M d'INDEL.

Si de nombreux variants génétiques sont aujourd'hui détectés grâce aux séquençages haut débit, il est également possible de localiser ceux impliqués dans la variation des caractères d'intérêt. La méthode utilisée est l'analyse d'association pangénomique dite GWAS (pour *Genome-Wide Association Studies*). Brièvement, le principe de la GWAS est d'étudier l'association entre de nombreux marqueurs génétiques polymorphes (généralement des SNP) bien répartis sur le génome et la variation d'un phénotype d'intérêt, et ce, pour un nombre d'individus le plus conséquent possible qui doivent donc être phénotypés pour le caractère d'intérêt et génotypés à ces multiples marqueurs. Pour chaque marqueur, l'effet sur la variation du phénotype d'intérêt est statistiquement testé par régression linéaire en mettant en lien les trois génotypes observés et le phénotype. Les marqueurs dits significativement associés au phénotype permettent de localiser des régions du génome impliquées dans la variation du caractère, ces régions sont nommées QTL (pour *Quantitative Trait Loci*) [393]. Notons qu'il est ici fait mention de marqueurs et non pas de variants causaux. En effet, il est extrêmement rare que le SNP marqueur soit aussi causal, les SNP identifiés sont fréquemment en déséquilibre de liaison (noté LD pour *Linkage Desequilibrium*) avec le variant causal qui lui a un effet réel sur le caractère. Ce variant causal peut être un autre SNP, un INDEL ou toute autre chose. Illustrons cette approche GWAS par quelques travaux emblématiques. Une des plus grandes études publiée à ce jour sur l'indice de masse corporelle chez l'humain et comptabilisant un total de 339 224 individus analysés sur 2,5M de marqueurs a mis en évidence 97 SNP quasi indépendants (soit 97 QTL) expliquant environ 2,7 % de la variance phénotypique [394]. Une autre étude du même ordre de grandeur composée de 253 288 individus analysés pour 2,5M de marqueurs et portant sur la taille chez l'humain a, quant à elle, permis d'identifier 697 marqueurs expliquant 16 % de la variance phénotypique [395]. La communauté scientifique ne s'attendait pas à ce que, pour des caractères connus comme héritables (>50 %) [396], l'ensemble des SNP/QTL détectés explique une si faible proportion de variabilité génétique. Cette observation, qualifiée « d'héritabilité manquante » par Manolio et al., [397, 398] est en partie expliquée par les effets trop faibles des milliers voir millions de SNP causaux non identifiés, le nombre d'individus analysés étant trop limité pour mettre en évidence des variants à si petits effets. Ainsi, pour les études évoquées en amont, il apparait que la prise en compte de tous les SNP, incluant les plus communs en termes de fréquence allélique, permet d'augmenter la part de la variance phénotypique expliquée aux alentours de 50 % pour la taille [395, 399] et de 20 % pour la

masse corporelle. Notons que d'autres facteurs tels que l'influence des réarrangements chromosomiques ou les effets GxE, encore mal connus, sont également à considérer.

Au vu de ces observations concernant l'héritabilité manquante, de la répartition sur l'ensemble du génome des signaux GWAS, de leur localisation à proximité de nombreux gènes sans lien évident avec le phénotype complexe et en général dans des régions non codantes, Boyle et al. ont proposé en 2017 le concept de « modèle omnigénique » [400]. Les auteurs suggèrent ainsi que les réseaux de régulation génétique sont suffisamment interconnectés pour que la quasi-majorité des gènes exprimés dans un tissu soit plus ou moins directement responsable de la régulation de l'expression des principaux gènes, en général codants des protéines, en lien avec le phénotype. Ainsi, la majeure partie de l'héritabilité pourrait être expliquée par les effets cumulés de ces nombreux gènes régulateurs encore inconnus et donc en dehors des voies biologiques connues pour être en lien avec le caractère.

3.1.3. Variants d'intérêt, où êtes-vous ? Des régions régulatrices plus ou moins à distance des gènes régulés

En lien avec ces observations, la communauté scientifique s'accorde aujourd'hui à dire que la vaste majorité des variants responsables de la variation des caractères complexes agissent par régulation de l'expression génique et sont donc dépourvus d'effets sur les structures protéiques tel qu'imaginé il y a quelques décennies. Cette hypothèse est un des piliers du modèle omnigénique de Boyle et al. indiqué plus haut et plusieurs études viennent supporter cette hypothèse. Comme montré par Pickrell et al. [401], seuls 2 % à 20 % des SNP seraient localisés dans des régions codantes, ainsi la majorité des polymorphismes associés à des phénotypes quantitatifs complexes se situeraient donc dans les régions non codantes du génome. De même, une étude de Maurano et al. indique que 76 % des SNP identifiés par GWAS se retrouvent dans les zones non codantes et plus précisément dans les régions hypersensibles à la DNase I pour 57 % d'entre eux [402]. Ces zones correspondent à des régions ouvertes de la chromatine facilitant l'accès à des facteurs extérieurs tels que des facteurs de transcriptions ou des éléments régulateurs non-codants (*e.g.*, les lncRNA, voir §1.2) et peuvent donc être associées des régions régulatrices de l'expression des gènes [403, 404]. Le consortium GTEx humain, dont l'objectif principal est d'étudier le lien entre variations

génétiques et variations d'expression, reporte des résultats de même nature [389]. À l'aide de 17 832 échantillons représentant 52 tissus, ce consortium souligne que les variants régulant localement, et donc pour la grande partie de manière directe (voir §3.2), l'expression des gènes sont à plus de 80 % dans des régions non codantes. Plus précisément, il estime à plus de 60 % la proportion de variants présents dans des régions introniques et plus de 20 % ceux dans des régions promotrices ou encore au niveau des UTR [389]. Les consortiums équivalents chez la poule (ChickenGTEx, voir Résultats §3.1) [405] et le bovin (CattleGTEx) [406], indiquent des proportions semblables. Ils montrent également que les variants régulateurs sont enrichis dans les zones promotrices et les *enhancer*. De manière intéressante, le variant régulateur (nommé *eVariant*) et le gène régulé associé (*eGene*) sont significativement enrichis dans les boucles de chromatine, en particulier les TAD (pour *Topological Associated domain*). Ces boucles permettent de rapprocher certaines régions du génome éloignées sur un plan linéaire, (*e.g.*, régions régulatrices et promoteurs d'un gène), permettant ainsi de moduler l'expression d'un gène [407]. En effet, en examinant les boucles 3D de la chromatine, le ChickenGTEx a dénombré 41 à 73 % de paires *eVariant-eGene* se situant dans des TAD qui ont été identifiés dans 22 tissues d'une étude annexe. Ce résultat est aussi observé pour le CattleGTEx qui montre par exemple que le *eGene* APCS et son *eVariant* situé à 144 kb en amont du site d'initiation de la transcription (TSS, pour *Transcription Start Site*) sont localisés dans un TAD et liés par une interaction chromatidienne significative. Notons que chez l'humain, la longueur moyenne d'un TAD est d'environ 880 kb avec, à l'intérieur de ces structures, des sous-TAD d'environ 185 kb [408]. Chez la poule, plusieurs études montrent respectivement des TAD d'une longueur moyenne comprise entre 148 kb et de 400 kb, sans distinction entre macro et micro-chromosomes [323, 409].

3.2. L'expression des gènes, de nombreux phénotypes à analyser par GWAS

Comme évoqué précédemment, la majorité des SNP marqueurs détectés par GWAS pour des caractères complexes sont localisés dans des séquences régulatrices agissant sur l'expression de gènes. Cette dernière est donc au cœur de la relation génotype-phénotype et son étude peut faciliter l'identification de zones du génome (voir de gènes) responsables de caractères plus complexes. D'un point de vue théorique, l'expression génique peut en conséquence être considérée comme un phénotype à part entière : c'est un caractère quantitatif dont la variabilité peut être expliquée par de la variabilité génétique et par des effets environnementaux. Il est de ce fait possible d'étudier sa composante génétique par GWAS et ainsi de localiser des régions dites, eQTL (pour *expression Quantitative Trait Loci*), responsable du caractère d'intérêt, ici étant l'expression du gène d'intérêt. Cependant, la réalité pratique et computationnelle apparait plus complexe. Alors que pour les caractères macroscopiques, il est d'usage de confronter quelques caractères d'intérêt (quelques unités à quelques dizaines) à des milliers de SNP, le nombre de phénotypes d'intérêt grimpe en flèche lorsqu'il s'agit d'expression génique. En effet, entre 10 000 et 15 000 PCG sont exprimés quel que soit le tissu considéré et ce nombre peut même atteindre 30 000 si l'on considère les lncRNA exprimés et certains tissus très riches en gènes exprimés comme le testicule (voir Résultats §1.1 et §1.2) [124, 125]. De plus, s'il est d'usage d'utiliser des puces de génotypages à basse/moyenne densité composées de quelques milliers de marqueurs pour des raisons économiques, certaines études utilisent des puces haute densité (*e.g.*, puces « Affymetrix Genome-Wide Human Array 6.0 » pour l'humain avec 906 600 marqueurs [410] ; puces « Affymetrix® Axiom® HD » pour la poule avec 580 954 marqueurs [411]) faisant grimper le nombre de marqueurs. De même, le RNAseq permettant de quantifier à la fois l'expression des gènes et d'extraire des génotypes fiables (voir §2.1 et Résultats §2.1), l'utilisation des centaines de milliers (voir millions) de marqueurs générés est également envisageable. Ainsi, face au nombre ubuesque de tests d'associations à réaliser (*e.g.,* 50 000 marqueurs et 20 000 phénotypes correspondant déjà à 10^9 tests), les analyses GWAS sur les gènes (nommées souvent analyses *eQTL*) exigent de corriger les résultats pour ces multiples tests. Par ailleurs, ces analyses nécessitent aussi une certaine puissance de calcul. En effet, alors que l'analyse GWAS de phénotypes classiques étaient réalisés sur processeurs CPU (pour *Central Processing Unit*), les ressources ne suffisent

plus et l'usage de processeurs graphiques GPU (pour *Graphical Processing Unit*) s'avère nécessaire permettant ainsi par leur structure de traiter un nombre plus important de calculs en parallèle [412]. Enfin, plusieurs logiciels dédiés à l'analyse eQTL et utilisant des astuces de calculs ont vu le jour tels que MatrixEqtl [413], FastQTL [414] et plus récemment TensorQTL [415]. C'est ce dernier qui est notamment utilisé dans le cadre des études eQTL réalisées à grande échelle par les consortiums GTEx chez l'humain ou la poule.

Au vu de la difficulté à traiter l'ensemble des paires variants-gènes et au vu des observations faites précédemment stipulant que les paires *eVariant-eGene* ont tendance à être regroupées dans des structures de taille inférieure à 1 Mb, la majorité des études s'intéresse aux marqueurs proches (généralement à 1 Mb) du gène considéré. Les régions eQTL ainsi détectées, proches du gène, et donc potentiellement à effet direct, ont traditionnellement des effets plus importants que les eQTL à distance [389]. Elles nécessitent donc un nombre d'échantillons raisonnable (une centaine d'individus) pour être détectés, ce qui n'est pas le cas pour la recherche de eQTL plus distants dont les effets sont en général indirects et de ce fait beaucoup plus faibles. Cette relation ambiguë entre la distance du variant et son effet direct ou indirect sur l'expression du gène est par ailleurs à l'origine d'une sémantique confuse :

- D'un point de vue positionnel :
    - Si le *eVariant* est situé à proximité du *eGene* (généralement moins de 1 Mb du TSS du *eGene*), le *eQTL* est qualifié de *local-eQTL*.
    - Si le *eVariant* est situé à distance du *eGene* (généralement plus de 1 Mb du TSS du *eGene*) ou sur un autre chromosome, le *eQTL* est qualifié de *distant-eQTL*.

- D'un point de vue mécanistique :
    - Si le *eVariant* a un effet direct sur l'expression du *eGene*, en étant dans une région régulatrice du gène, par exemple dans un intron, un promoteur ou encore un *enhancer*/*silencer*, il est qualifié de *cis-eQTL*.
    - Si le *eVariant* a un effet indirect sur l'expression du *eGene*, notamment via une molécule intermédiaire tel qu'un facteur de transcription ou un lncRNA, alors il est considéré comme *trans-eQTL*.

Ainsi, s'il est courant dans la littérature de voir le terme *cis*-eQTL employé pour qualifier des *local*-eQTL avec un effet *cis* fortement supposé, il serait plus rigoureux de conserver le terme de *local*-eQTL en attendant des éléments supplémentaires attestant de la réelle nature *cis* du

eVariant. En effet, bien que cette relation entre *cis* et *local* semblent généralement vérifiée, un gène étant caractérisé par de multiples séquences *cis* de régulation, certains travaux rapportent des lncRNA à proximité du PCG (voir Résultats §1.1 et §1.2) et qui agirait comme régulateur de l'expression du PCG [124, 125, 416]. Un variant associé à l'expression du PCG, et qui agirait via son impact sur la structure ou l'expression d'un lncRNA localisé à sa proximité, serait alors défini comme un *trans*-*local*-eQTL pour le PCG et non comme un cis-eQTL. Notons que si la vision mécanistique nécessite des manipulations de biologie moléculaire et cellulaire ou des analyses complémentaires sur l'expression (*e.g.*, par étude ASE) pour déterminer la nature de la régulation, la vision positionnelle ne fait, quant à elle, aucun a priori si ce n'est de fixer un seuil de distance pour distinguer le *local* du *distant*.

3.3. Mettre en parallèle les eQTL avec les QTL : l'expression génique un phénotype intermédiaire entre génotype et phénotype

Comme vu précédemment, des centaines, voire milliers, de variants génétiques ont pu être associées par GWAS à divers phénotypes d'intérêt dans de nombreuses espèces (*e.g.*, dbQTL animal [417]). Cependant, la grande majorité des variants détectés se trouvent dans des régions non codantes du génome et leurs actions restent encore compliquées à identifier. Une approche pour aider à leur identification consiste à rechercher le gène régulé, en réalisant conjointement des études GWAS de type eQTL sur l'expression des gènes et QTL sur les caractères complexes d'intérêt. Cette approche repose sur la nature des variants recherchés qui sont des variants régulateurs de l'expression génique. Dans ce contexte, il convient alors d'analyser si les deux signaux GWAS générés par les deux études d'associations, correspondent à un même variant causal partagée. Les premières approches ont consisté en une comparaison visuelle des chevauchements des signaux GWAS eQTL et QTL, mais l'abondance des eQTL rend très probable un chevauchement accidentel entre signaux.

Ainsi, d'autres approches, elles statistiques, ont été développées pour tester la colocalisation entre eQTL et QTL. L'une des premières méthodes apparues est COLOC [418] qui calcule des probabilités selon les modèles suivant : *absence d'association (H0), association avec l'expression du gène uniquement (H1), association avec le phénotype uniquement (H2), associations distinctes pour l'expression du gène et le phénotype (H3), association commune aux deux (H4)*. Un odds ratio élevé

en faveur de H4 par rapport aux autres modèles indique alors une colocalisation probable. D'autres méthodes comme eCAVIAR [419] ou ENLOC [420] estiment la probabilité de colocalisation en considérant le LD (entre les variants des 2 signaux GWAS, LD que l'on suppose élevé s'il s'agit du même variant causal. Notons que cette approche peut également être appliquée pour des paires d'eQTL pour observer des phénomènes de co-régulation mais également pour des phénotypes macroscopiques.

Une hypothèse biologique importante derrière ces analyses conjointes eQTL/QTL est que les eQTL responsables d'une part du caractère d'intérêt doivent être analysés dans un tissu ayant un lien supposé avec le phénotype d'intérêt. S'il n'y a pas d'apriori sur le tissu, une alternative est d'utiliser plusieurs tissus, ce qui diminue d'autant le nombre d'individus analysés et donc le pouvoir statistique de l'étude [389], l'acquisition d'un transcriptome étant coûteuse. De plus, certains gènes candidats peuvent être manqués, par exemple si les eQTL sont spécifiques d'un type cellulaire non étudié. Plusieurs études récentes ont appliqué ces approches pour mettre en évidence des gènes candidats impliqués dans divers phénotypes complexes. Beesley et al., ont par exemple utilisé ces analyses de colocalisation pour identifier des *eGene* et variants sous-jacents en lien avec le risque de développer un cancer du sein chez l'humain. Dix-sept gènes ont été considérés comme médiateurs potentiels. Pour l'un d'entre-eux, NTN4, un *eVariant* était situé dans un élément *enhancer* qui interagissait physiquement avec le promoteur du gène et réduisait notamment son activité. Fort de ce constat, les auteurs ont réalisé un *knockdown* de NTN4 dans les cellules mammaires et une augmentation de la prolifération cellulaire *in vitro* et de la croissance tumorale *in vivo* ont été observées [421].
En conclusion, l'analyse de la colocalisation ne suffit pas à prouver une relation de causalité entre l'expression du gène et le phénotype. Des analyses expérimentales complémentaires sont nécessaires pour confirmer le rôle des gènes candidats. Malgré cela, la colocalisation peut fournir des pistes prometteuses pour hiérarchiser les gènes et variants candidats causaux en lien avec des phénotypes d'intérêt complexes.

Résultats

- Articles et travaux complémentaires -

** Les résumés s'appuient sur les figures de l'article associé, nécessitant donc de s'y référer selon les indications fournies.*

*** Les numéros de référence correspondent à ceux utilisés dans le cadre du présent manuscrit et non à ceux des articles publiés.*

1. Annotation des gènes du génome

1.1. Production d'un atlas enrichi en gènes de type lncRNA et PCG pour l'assemblage GRCg7b et annotations fonctionnelles à travers 47 tissus *(Résumé d'article)*

1.1.1. Contexte et objectifs

La caractérisation du contenu génique des chromosomes d'un organisme, *i.e.*, des régions transcrites, est cruciale pour la plupart des études génétiques, notamment pour identifier les gènes et variants génétiques associées à des phénotypes d'intérêt. Cependant, alors que les gènes codant des protéines (PCG) sont relativement bien décrits dans les génomes, les loci géniques associés aux ARN longs non-codants (lncRNA) sont encore mal décrits [99]. Les lncRNA, découvert massivement dans le génome humain au début des années 2010, sont des régulateurs de l'expression des gènes via divers mécanismes [187]. Ils sont particulièrement impliqués dans la régulation de la structure de la chromatine, la transcription, l'épissage des ARN, la stabilité des ARN et la traduction [188]. Ils participent à divers processus biologiques au niveau cellulaire et de l'organisme, influençant donc les phénotypes observés en accord avec le modèle polygénique additif. Par conséquent, une cartographie complète des régions transcrites codantes et non-codantes est nécessaire pour comprendre les relations génotype-phénotype. Les annotations de référence du génome humain (GRCh38.p13) et de la souris (GRCm39) par Ensembl (février 2023, v109) comprennent respectivement 18 882 et 11 621 lncRNA [422, 423]. Ces nombres de lncRNA connus sont amenés à augmenter avec les efforts de recherche [113, 424]. Pour les espèces d'élevage et plus précisément pour la poule, les lncRNA sont de plus en plus intégrés dans les annotations de référence du génome, mais ces dernières restent encore très incomplètes, n'atteignant pas les nombres identifiés chez les espèces modèles et montrant des divergences entre annotations en termes de modèles de transcrits et de gènes. Notons que, pour les PCG, si les modèles de gènes sont assez communs entre annotations, une variabilité assez importante demeure au niveau des modèles de transcrits [114]. Les caractéristiques spécifiques des lncRNA (faible expression, forte spécificité tissulaire et conditionnelle…) et le nombre limité d'échantillons d'ARN utilisés pour générer ces annotations, la plupart du temps issus de données *short-read*, expliquent en partie ces observations. Cependant, l'évolution des techniques de séquençage et la diminution de leur coût, notamment pour le *long-read* et ses alternatives [425], apparait à

court termes comme une solution à l'amélioration de ces annotations, comme cela a déjà été fait chez la souris et l'humain par le consortium GENCODE [426].

Afin d'améliorer l'exhaustivité de l'annotation du génome, en particulier pour les lncRNA, une stratégie consiste à combiner les annotations de référence "RefSeq", "Ensembl" et d'autres bases de données supplémentaires. Un atlas enrichi en gène et notamment en lncRNA pour la poule et intégrant notamment les modèles de gènes de "Ensembl", "RefSeq" et d'autres bases de données [124] avait déjà été proposé en 2020. Cependant, depuis, la nouvelle version d'assemblage du génome de poulet GRCg7b et ses annotations de référence associées ont été publiées, conduisant à mettre à jour, mais également à améliorer cet atlas de gènes. Ainsi, de nouvelles données telles que des ressources multi-tissus de FAANG [195, 196] ont été apportées, conduisant au total à l'identification de 24 102 PCG et 44 428 lncRNA. D'autre part, une annotation fonctionnelle a été produite en utilisant 1400 échantillons RNAseq provenant de 47 tissus ou types cellulaires.

Ainsi, les profils d'expression des gènes ont été caractérisés et mis en forme pour faciliter l'extraction du ou des tissus dans lesquels un gène d'intérêt est le plus exprimé, afin d'orienter les hypothèses fonctionnelles et les études expérimentales. De plus, en supposant qu'un gène exprimé dans un tissu ou groupe de tissus joue un rôle lié aux fonctions de ce tissu [427], une analyse approfondie de la spécificité d'expression tissulaire des lncRNA et des PCG a été réalisée. Les configurations entre les gènes et notamment entre les PCG, lncRNA et miRNA ont été caractérisées. Des cas intéressants de lncRNA et PCG hébergeant des miRNA ou encore de lncRNA antisense de PCG, conservés chez l'humain et pour lesquels les profils d'expression entre la poule et l'humain étaient similaires et cohérents, suggérant une fonction commune [428], ont été identifiés.

1.1.2. Résultats

Six bases de données contenant à la fois des modèles de PCG et de lncRNA ont donc été sélectionnées pour créer une annotation du génome enrichie (voir Figure 6). Cela inclut les bases de référence RefSeq et Ensembl, ainsi que des annotations de projets multi-tissus comme Fr-AgENCODE, UC Davis, l'annotation INRAE et la base Noncode dédiée aux ARN non-codants. La comparaison du contenu des modèles de gènes entre les bases montre un chevauchement plus important des PCG que pour les lncRNA (voir Figure 1A). En effet, si les taux de chevauchement atteignent environ 95 % pour les PCG (hors cas spécifiques), ceux des lncRNA dépassent exceptionnellement les 50 % mais sont plus souvent de l'ordre de la dizaine de pourcents. Notons que pour les bases de données de référence (RefSeq et Ensembl) ce taux est de 37 %. Ces faibles pourcentages de chevauchement des lncRNA, mais aussi le taux inférieur à 100 % pour les PCG justifient à eux seuls l'agrégation de ces ressources. Ces variabilités de recouvrement peuvent en partie s'expliquer par un nombre variable de modèles géniques, notamment pour les lncRNA (5 789 pour RefSeq et 11 944 pour Ensembl, par exemple), et par des modèles géniques différemment caractérisées (voir Figure 1B). La concordance des TSS avec les pics CAGE du projet FANTOM est meilleure pour les PCG, variant entre 60 et 40 %, contre un maximum de 15 % pour les lncRNA (voir Figure 1C). Sur la base de ces observations, les annotations ont été intégrées en ajoutant successivement les loci de gènes de chaque base, sans chevauchement. L'ordre suivant a été choisi : RefSeq, Ensembl, Fr-AgENCODE, UC Davis, INRAE, Noncode. L'atlas enrichi contient finalement 78 323 modèles géniques incluant 24 102 PCG, 44 428 lncRNA et 991 miRNA (voir Figure 1D). À la suite de cette agrégation, la densité de PCG et de lncRNA est corrélée et plus élevée dans les micro-chromosomes, mieux représentés dans le dernier assemblage GRCg7b comparé au précédent GRCg6a (voir Figure 1E).

Parmi les 78 323 gènes, un total de 63 513 gènes (81 %) est considéré comme exprimé, incluant 22 468 PCG (93 %) et 35 257 lncRNA (79 %) au travers des 47 tissus représentés (voir Figure 2A et 2B). Le nombre de gènes exprimés selon les sources d'annotation avoisinent les 75 % avec un minimum de 49 % pour Noncode et un maximum de 91 % pour RefSeq. Les ACP sur les données d'expression, indépendamment du type de gène, ont résulté en un regroupement dépendant du tissu et non du projet, validant la consistance de nos données

d’expression et en particulier l’étape de normalisation (voir Figure 2C et D). Notons que la normalisation TPM est plus efficace que le TMM dans ce cadre d’étude. Il est cependant à noter que les lncRNA sont d'abord regroupés selon les tissus avec le plus de gènes spécifiques tels que les testicules, le cerveau ou les tissus/cellules en lien avec l’immunité.
De manière plus précise, le nombre de gènes exprimés dépend du type de gène et du seuil d'expression : 88 % des PCG ont une expression ≥ 1 TPM dans au moins un tissu contre 57 % des lncRNA. Pour un seuil à 0,1 TPM, le nombre de PCG exprimés varie de 9 887 (44 %) à 17 747 (79 %) avec une moyenne de 14 837 (66 %) par tissu. 7 485 PCG sont exprimés dans tous les tissus. Pour les lncRNA, le nombre exprimé varie de 1 189 (3 %) à 16 708 (47 %) avec une moyenne de 7 646 (21 %) par tissu. Seulement 103 lncRNA sont exprimés dans tous les tissus. Avec un seuil à 1 TPM, le nombre moyen chute à 11 139 PCG et 1 972 lncRNA, cohérent avec le fait que les lncRNA sont connus pour être moins exprimés que les PCG.

La spécificité tissulaire, calculée par la valeur tau (τ), varie selon le seuil d'expression appliqué pour considérer un gène comme exprimé (voir Figure 3A). Avec un seuil à 0,1 TPM, 86 % des gènes sont spécifiques d'un tissu (TS) contre 46 % à 1 TPM. Le choix a alors été fait de travailler sur les gènes ayant une expression ≥ 1 TPM dans au moins un tissu, soit 20 252 lncRNA et 19 819 PCG. Les PCG et lncRNA ont alors des distributions de valeurs de tissu-spécificités différentes, les lncRNA étant globalement plus TS que les PCG. En effet, 23 % des PCG ont un $\tau \geq 0,9$ contre 68 % pour les lncRNA. De manière empirique, il a été observé que les gènes considérés comme TS par la mesure du tau, peuvent en réalité être exprimés dans plusieurs tissus. Selon la classification présentée en « Mat. et Met. », environ 72 % des gènes sont *mono-TS*, 23 % sont *poly2to7-TS* et 5 % sont *poly8to47-TS*, cette distribution étant identique entre PCG et lncRNA (voir Figure 3B). De manière plus précise, il apparait que la proportion de gènes TS dans chaque catégorie est très variable selon les tissus (voir Figure 3C). Par exemple, 74 % des gènes TS du testicule sont *mono-TS* alors que seulement 0,8 % des gènes TS du duodénum le sont. Cette variabilité de la proportion est liée aux autres tissus présents dans l'ensemble de données et à leurs fonctions communes, comme pour les tissus associés au système intestinal ou au système cérébral. Inversement à cette observation, il est à noter qu’un gène peut être fortement exprimé dans un tissu sans être TS.

Afin de tester le lien entre expression et phénotype, les 54 traits mendéliens identifiés par OMIA chez la poule ont été analysés. Parmi les 36 traits qui peuvent être associées de manière forte à un tissu, environ 60 % (n = 17) ont un gène causal dont l'expression est cohérente avec le tissu attendu et présent dans notre jeu de données. Plusieurs exemples sont fournis (voir Figure 4A) et notamment le cas de SV2A (*synaptic vesicle glycoprotein 2A*), gène codant une protéine impliquée dans les tissus cérébraux (voir Figure 4B). Ce gène était initialement connu en galgal2 mais a disparu des bases de données jusqu'à son retour à partir de galgal5 et jusque dans la dernière version GRCg7b. Ainsi, même si son expression n'est pas spécifique à un tissu ($\tau = 0,81$), il est fortement exprimé dans le cortex, le cerveau, l'hypothalamus et le cervelet, a l'image de son orthologue humain. Cependant, certains gènes méritent une analyse complémentaire poussée, comme le gène SLCO1B3 impliqué dans la couleur bleutée des œufs et attendu comme impliqué au niveau de l'utérus, mais dont l'expression est ici spécifique du foie, comme son orthologue humain.

Les expressions ont également été analysées sous l'angle des différentiels d'expressions (DEG) entre sexes où 4 206 gènes au total ont été identifiés au travers de six tissues avec au moins huit individus de chaque sexe provenant du même jeu de données (voir Figure 4C et D). Le nombre de DEG repérés varie de 2 475 pour le foie à 233 gènes pour la glande de Harder. Les gènes correspondent à 816 lncRNA, 3 276 PCG (soit 8,3 % et 19,9 % du total de lncRNA et PCG exprimés) et 114 autres types de gènes. Parmi eux, 3 384 (80,5 %) sont identifiées comme DEG dans un seul tissu, avec des pourcentages similaires pour les lncRNA (85,9 %) et les PCG (79,5 %). Notons que la majorité (84,1%) des DEG dans deux tissus ou plus ont des directions de *fold-change* cohérentes entre tissus. Comme attendu, un enrichissement des gènes du chromosome Z (821 gènes, 19,5 %) est observé alors qu'ils ne représentent que 5 % du total des gènes exprimés. La médiane de *log(fold-change "mâle/femelle")* de 0,76, reflète par ailleurs la compensation incomplète de dosage des chromosomes sexuels connue chez la poule.

Les 991 miRNA identifiés dans l'annotation enrichie ont été classifiés positionnellement par rapport au lncRNA ou PCG le plus proche. 244 (24,6 %) et 717 (72,4%) miRNA sont situés respectivement dans un intron ou un exon de 194 lncRNA et 627 PCG. Plus précisément, 43,8 % des miRNA sont dans un intron et 51,6 % dans un exon d'un lncRNA contre 65,4 % et 32,8 % respectivement pour les PCG. En se concentrant sur les 179 lncRNA exprimés

(TPM ≥ 0,1) dans au moins un tissu, 133 (74,3%) ont une expression ≥ 1 TPM, une proportion significativement plus élevée que pour le total des lncRNA (74,3 % vs 56,3 %). La même tendance est observée pour les 622 PCG exprimés associés à 712 miRNA (98,2 % vs 87,8 %). De même, 110 (61,5 %) lncRNA sont spécifiques d'un tissu, avec une proportion similaire au total des lncRNA. Comme attendu, ce taux est plus élevé que pour les PCG dont seulement 61 (9,8 %) sont spécifiques d'un tissu. Certains lncRNA spécifiques d'un tissu hébergeant des miRNA sont conservés chez l'humain avec des profils d'expression cohérents (voir Figure 5A et 5B). Par exemple, LOC124417505 hébergeant MIR122-1 dans un exon est spécifique du foie comme son orthologue MIR122HG. D'autres lncRNA spécifiques de tissu et nouvellement modélisés semblent aussi orthologues à des lncRNA connus chez l'humain.

Afin d'hypothétiser sur des relations biologiques significatives entre les lncRNA et les PCG selon le principe du « *guilt-by-association* », les PCG et lncRNA ont été classés selon leur configuration avec le PCG le plus proche et les co-expressions entre les paires de gènes ont été calculées sur les 47 tissus (voir Figure 5C). Sur les 35 257 lncRNA et 22 468 PCG exprimés, 33 907 (94,4 %) et 20 656 (91,9 %) sont associés à un PCG dans une fenêtre de 1 Mb. Parmi eux, 2 331 (6,9 %) paires lncRNA:PCG et 3 375 (16,4 %) paires PCG:PCG montrent une co-expression positive significative avec une co-expression plus importante pour les paires PCG:PCG que pour celles des lncRNA:PCG ($|\rho| = 0,32$ vs 0,16). Cependant, si la co-expression peut permettre de poser des hypothèses sur la fonctionnalité d'un lncRNA, il convient tout de même de vérifier la véracité du modèle génique, notamment lorsque des données de séquençage *short-read* ont été utilisées. En effet, la longueur des *reads* couplée à une faible profondeur peuvent mener à des modélisations erronées de lncRNA sur le même brin (*same strand*) en amont/5'UTR ou en aval/3'UTR du PCG. Ainsi, les lncRNA en aval/3'UTR d'un PCG (12,6 %) apparaissent plus co-exprimés comparé aux autres configurations, spécialement les lncRNA en amont/5'UTR (5,2%). Des tests PCR ont confirmé trois paires comme extensions du PCG, et trois autres comme lncRNA indépendants. De plus, les lncRNA et PCG en configuration « *same strand* » et « *divergent* » avec un autre PCG montrent des valeurs de co-expression plus élevées qu'en configuration « *convergent* ». Notons également, qu'en excluant les paires en aval/3'UTR et en se focalisant sur les paires intergéniques, un enrichissement des gènes co-exprimés à une distance ≤ 5 kb comparé à une distance ≥ 5 kb pour les configurations « amont/5'UTR» et «divergent» est observé.

Au final, ce travail propose une annotation du génome (fichier *.gtf*) et des gènes (fichier *.tsv*) construite sur l'assemblage GRCg7b et intégrant les bases de données de référence Ensembl et RefSeq. Ce changement d'assemblage et sa coexistence avec les précédents (GRCg6a) et l'alternatif (GRCg7w) a conduit à des changements importants d'identifiants de gènes dans certaines bases, notamment Ensembl, compliquant la transition et créant des incertitudes entre études sur différents assemblages et annotations. Pour faciliter la comparaison entre études et assemblages, une table d'équivalence a été établie permettant le transfert d'une annotation à une autre.

1.1.3. Discussion et conclusion

Ce travail propose une solution pour enrichir l'annotation du génome et des gènes du génome de la poule tout en conservant les informations des deux bases de référence, "RefSeq" et "Ensembl". Bien que l'utilisation d'un *pipeline* de modélisation de gènes unique incluant toutes les données de séquençage brutes soit la meilleure solution, cette approche offre une bonne alternative, car *i)* elle unifie les deux annotations du génome les plus utilisées (comme le projet MANE – *Matched Annotation from NCBI and EMBL-EBI*) disponible actuellement uniquement pour l'humain [429] ; *ii)* elle conserve les identifiants "RefSeq" et "Ensembl" pour les loci de gènes communs ; *iii)* elle est plus rapide qu'une annotation *de novo*, et est adaptable aux changements majeurs dans les versions successives des bases de données de référence.

Cette approche augmente l'exhaustivité de l'annotation du génome de la poule, et en particulier pour les lncRNA, qui sont plus difficiles à identifier que les PCG en raison de leur faible expression spécifique aux tissus et conditions [103, 424, 430]. De plus, du fait de l'utilisation de données de séquençage *short-read*, les modèles de transcrits sont mal décrits, quels que soient les biotypes de gènes, même si cette tendance est plus marquée pour les lncRNA que pour les PCG [114]. À titre d'exemple, sur les six bases de données utilisées dans cette étude, le nombre médian maximal de transcrits par lncRNA et PCG était respectivement d'un et trois, des chiffres inférieurs à ceux observés chez l'humain, avec trois et sept transcrits en moyenne par lncRNA et PCG respectivement [103, 114]. D'autre part, le taux de chevauchement entre le TSS des transcrits et les pics CAGE, qui est loin de 100%, même pour les PCG, souligne une modélisation encore incomplète des transcrits. L'émergence et la démocratisation des technologies *long-read*, dont l'inconvénient aujourd'hui est la capacité à obtenir des profondeurs de séquençage comparables aux technologies *short-read* et limitant ainsi leur utilisation massive pour les études axées sur l'expression des gènes [431], permettront dans l'avenir de clarifier ces modèles.
La faible profondeur du *long-read* pourrait expliquer pourquoi la base de données "Davis", principalement basée sur ce type de séquençage, identifie des lncRNA principalement mono-exoniques et généralement situés sur le même brin des introns de PCG, cause alors d'un faible taux de chevauchement avec les pics CAGE. Une autre limitation est que certains loci de gènes peuvent être erronés, en particulier pour les lncRNA qui sont sur le même brin d'un PCG

proche, et fortement co-exprimés. Ces lncRNA pourraient dans la pratique être une région transcrite non traduite (UTR) du PCG qui sont, comme les lncRNA, difficiles à modéliser avec du *short-read* et nécessitent des analyses complémentaires [40, 41]. Par conséquent, une validation expérimentale, par PCR par exemple, est nécessaire pour vérifier l'existence de tels lncRNA avant d'analyser plus en détail leurs fonctions.

L'annotation de gènes basée sur l'expression dans 47 tissus souligne que 81 % des modèles de gènes étaient exprimés dans au moins un tissu, indiquant que ces modèles ne sont pas du bruit transcriptionnel. Comme indiqué dans la littérature pour les analyses inter-espèces, les lncRNA sont préférentiellement exprimés dans les organes sexuels tels que les testicules [276, 432] et dans un second temps par les tissus liés au cerveau [99, 433, 434]. Cette expression favorisée est potentiellement associée à un environnement chromatinien facilitant la transcription d'éléments putativement non fonctionnels et permettant l'émergence de nouveaux gènes. De même, en cohérence avec la littérature, une proportion de lncRNA tissus-spécifiques plus élevée par rapport aux PCG [99, 124] a été observée. La caractérisation des profils d'expression fournit des informations essentielles pour sélectionner les lignées cellulaires pertinentes permettant d'étudier par la suite les fonctions des gènes. Cela peut également être une première indication de sa fonction, en particulier pour les gènes spécifiques aux tissus, comme l'illustre l'analyse du profil d'expression des gènes causaux associés à des caractères mendéliens.

Cette étude souligne également la relativité de la mesure de la spécificité tissulaire, qui dépend de plusieurs facteurs, dont la métrique et la valeur seuil, mais surtout du nombre et type de tissus. L'ajout d'un autre tissu, avec ou non des fonctions similaires, peut grandement faire varier les valeurs de spécificité tissulaire des gènes, en particulier lorsque peu de tissus sont considérés. À titre d'exemple, en utilisant un panel de 21 tissus, l'étude sur l'atlas enrichie proposée en 2020 [124] montre un taux de spécificité tissulaire de 25 % pour les lncRNA contre 10 % pour les PCG, contre respectivement 68 % et 23 % observés dans cette étude. La métrique utilisée apparait alors déterminante.

Concernant les 4 206 gènes différentiellement exprimés (DEG) selon le sexe et observés dans six tissus, un pourcentage inférieur à celui rapporté par le consortium GTEx humain est

observé (19,8 % du total des PCG exprimés contre 37 % de tous les gènes chez l'humain), probablement en raison du nombre plus élevé de tissus analysés [435]. De manière intéressante, 80 % des DEG selon le sexe sont observés dans un seul tissu malgré une expression dans plusieurs d'entre eux, suggérant une régulation dépendante du tissu. Ces observations sont en accord avec les travaux de Oliva et coll., 2020 [435] même si ce pourcentage est surestimé dans notre étude en raison du faible nombre de tissus analysés. Certains gènes rapportés dans des études antérieures comme différentiellement exprimés entre les sexes chez les mammifères ont également été trouvés chez la poule tel que CYP3A4 lié au métabolisme des médicaments [435, 436], VWCE (alias urg11) prédit pour permettre l'activité de liaison aux ions calcium [435, 437], ou encore la polykystine 2 (PKD2), une protéine membranaire impliquée dans un canal cationique perméable au calcium [437].

D'autre part, les résultats indiquent que la plupart des 991 miRNA sont situés dans un gène, avec 75 % d'entre eux dans un PCG et 25 % dans un lncRNA. Ces résultats sont en accord avec ceux de Liu et al., 2018, [438] qui ont démontré qu'une grande fraction des miRNA présent dans miRBase V21 (1 325 sur 1 881) sont également hébergés dans un gène et avec ceux de Dhir et al., 2015, [428] qui ont rapporté, chez l'humain, une petite fraction de miRNA (17,5 %) hébergés par un lncRNA. La localisation des miRNA par rapport au gène le plus proche est un facteur important à considérer pour étudier la régulation transcriptionnelle des miRNA primaires, qui n'est pas encore totalement comprise. Des études antérieures chez l'humain ont rapporté que plus de la moitié des miRNA résident dans des introns de PCG (aucune étude ne s'intéressant spécifiquement aux lncRNA) et sont supposés être co-exprimés avec leurs gènes hôtes, dérivant de transcrits primaires communs [439–442]. Cette hypothèse doit être nuancée puisque Ozsolak et al., 2008 [443], ont rapporté qu'une fraction significative de miRNA intra-géniques étaient initiés indépendamment des transcrits PCG. Notons que les noms des lncRNA de poule hébergeant des miRNA ne sont pas normalisés et devraient s'appeler MIRxxxHG comme MIR155HG, le seul lncRNA correctement nommé, à la suite du précédent travail publié en 2020 [124].

Les analyses des configurations lncRNA:PCG montrent que les lncRNA ont tendance à être plus géniques qu'intergéniques. Cette observation semble contradictoire avec la littérature [99, 444], où l'on souligne *i)* l'utilisation de données RNAseq non orientées pour les publications

les plus anciennes ainsi que *ii)* la prise en compte uniquement des modèles de transcrits multi-exoniques par les *pipelines* bio-informatiques pour éviter les faux positifs potentiels correspondant à des transcrits faiblement couverts. De plus, la baisse du coût du RNAseq *short-read* permet maintenant de séquencer avec une plus grande profondeur et de mieux considérer les transcrits moindrement exprimés qu'ils soient intergéniques et/ou mono-exoniques. Dans notre étude, une surévaluation des lncRNA « *same-strand* » a été observée et pourrait être expliquée par l'utilisation d'une base de données de séquençage *long-read*, limitée en profondeur, comme mentionné précédemment. Les lncRNA impliqués dans de telles configurations doivent être considérés avec précaution, puisque, comme illustré dans le manuscrit, certains d'entre eux sont en réalité le reliquat d'un PCG mal modélisé. En effet, de nombreux isoformes de PCG sont encore mal annotés, en particulier pour les espèces non modèles. Par exemple, comme le montrent Lagarrigue et al., 2021 [103], pour un nombre stable de modèles de gènes, le nombre de transcrits PCG oscille entre 28 000 et 50 000 pour les espèces d'élevage tandis qu'il dépasse 100 000 pour la souris et 150 000 pour l'humain. Une valeur de co-expression très élevée dans les tissus (ou intra-tissu selon l'étude) et une faible distance entre les modèles de gènes peuvent être considérées comme un indicateur de méfiance. Par exemple, Muret et al., 2019 [160] ont montré par une validation PCR que le lncRNA FLRL7 en « *same-strand* » de FADS2 dans la souris constituait en réalité un seul modèle de gène. Cependant, si certaines paires lncRNA:PCG en « *same-strand* » doivent être considérées avec précaution, une partie considérable des lncRNA constitutifs semblent exister de manière indépendante et il est ainsi possible de proposer des hypothèses concernant la fonction de lncRNA en appliquant le principe de « *guilt-by-association* » [445].

1.1.4. Matériels et démarches

Création de l'atlas enrichi en modèle génique – Ces travaux ont été réalisés avec l'assemblage de référence bGalGal1.mat.broiler.GRCg7b (GCF_016699485.2) de la poule [446]. Cet atlas est issu de la combinaison et la priorisation de six sources différentes incluant :

- Les modèles de gènes, selon l'assemblage GRCg7b, tels que fournis par les annotations de référence RefSeq (v106) [447] et Ensembl (v107 – qui intègre les données du projet GENESWitCH) [448].
- Les modèles de gènes issus des projets pilotes multi-tissus FAANG (*Functional Annotation of Animal Genomes*) [196] selon l'assemblage GRCg6a : le projet FR-AgENCODE [323] sur 11 tissus (2 mâles et 2 femelles par tissu) et le projet FarmENCODE [372] sur 15 tissus (1 mâle et 1 femelle par tissu). Notons que le projet FarmENCODE inclut des échantillons séquencés avec la technologie de *long-read* Nanopore d'Oxford.
- Les modèles de gènes modélisés dans la version précédente de l'atlas par Jehl et al., 2020 [124] selon l'assemblage galgal5.
- Les modèles de gènes provenant de la base de données NONCODE v6.0 [112] ne contenant que des modèles de gènes non-codants issus de la littérature et de bases de données publiques, selon l'assemblage galgal4.

Pour les annotations basées sur un assemblage antérieur à GRCg7b, les coordonnées des modèles géniques selon cet assemblage ont été obtenues via le service de *remapping* de génome (*Coordinate remapping service*) du NCBI [449]. Les pics CAGE (*Cap Analysis of Gene Expression*) robustes du projet FANTOM5 [385] ont été converties de l'assemblage galgal5 à GRCg7b par *remapping* [449]. Par la suite, un transcrit était considéré comme bien modélisé si son TSS (*Transcription Start Site*) en 5' chevauchait un pic CAGE dans une fenêtre de +/- 30 pb.

En considérant, pour chaque base de données, leur qualité intrinsèque, la concordance de leurs modèles géniques avec les pics CAGE et leur popularité, les six sources ont alors été ajoutées successivement dans l'ordre suivant (voir Figure 6A) : 1) RefSeq ; 2) Ensembl ; 3) FR-AgENCODE ; 4) Davis ; 5) Inrae ; 6) Noncode. Ainsi, un modèle génique et les transcrits affiliés provenant de la source N+1 n'étaient ajoutés que s'ils ne chevauchaient pas des modèles préalablement existants et provenant des sources 1 à N. Deux modèles étaient ainsi considérés comme se chevauchant si au moins un de leurs transcrits avait au moins un exon commun sur le même brin avec un recouvrement d'au moins 1 pb (voir Figure 6B). La détection des chevauchements a été réalisée avec l'outil BEDTools [450]. Pour améliorer l'ajout successif

des différents modèles, une décomposition par classe de biotype a été utilisée. Cette approche a permis de limiter les recouvrements de modèles similaires, mais avec des biotypes différents ou non assignés.

Création de l'annotation fonctionnelle des gènes – L'expression des gènes a été quantifiée à partir de 36 jeux de données publiques pour un total de 1400 échantillons et 47 tissus et/ou modèles cellulaires représentant la diversité des systèmes physiologiques de la poule. Les séquences FASTQ ont été alignées sur le génome de référence GRCg7b et l'expression a été quantifiée selon l'annotation enrichie, en utilisant le *pipeline* d'analyse nf-core « *rnaseq* » (v3.8.1) [451, 452] fournissant les expressions en *counts* et TPM. Pour chaque tissu dans chaque projet, une médiane des expressions en TPM sur tous les échantillons a été calculée. Pour les tissus présents dans plusieurs projets, la médiane a été calculée à partir des médianes précédemment obtenues dans chaque projet. Un gène était ainsi considéré comme exprimé s'il remplissait les critères suivants :

- Expression médiane ≥ 0,1 TPM dans au moins un tissu
- Au moins 50 % des échantillons d'un tissu dans un projet donné avec un nombre de *reads* ≥ 6, et expressions TPM et TMM normalisées ≥ 0,1.

La normalisation TMM (*trimmed mean of M-values*) a été réalisée via le package edgeR (v3.32.1) [329, 358] à partir des *counts* bruts.

Afin d'observer la distribution des tissus et projets selon les données d'expressions, des ACP ont été réalisées avec le package FactoMineR (v2.7) [453] sur les données d'expressions pour les gènes exprimés et transformées selon le log2(TPM+1) pour les 1400 échantillons. À partir des médianes d'expression par tissu, un dendrogramme basé sur la matrice de distance calculée avec (1 - corrélation de Pearson) des expressions en log2(TPM+1), et une classification hiérarchique faite par la méthode de Ward ont été réalisés.

Une étude de la spécificité tissulaire a été réalisée en utilisant la métrique usuelle tau (τ) [378] [82] avec l'expression médiane des tissus transformée en log10. Cette dernière varie de 0 pour une expression identique dans tous les tissus à 1 pour une expression spécifique dans un seul et unique tissu. Un gène était alors considéré comme spécifique de tissus pour $\tau \geq 0{,}90$ et dans certaines analyses avec un filtre sur l'expression (≥ 1 TPM dans au moins un tissu). Ces gènes

considérés comme spécifiques ont été classés en trois catégories selon leur profil d'expression et la présence d'un écart d'expression d'un facteur ≥ 2 entre les expressions consécutives de deux tissus, on distingue alors les gènes spécifiques :

- d'un tissu unique (notés *mono*_TS) ;
- d'un groupe de 2 à 7 tissus (notés *poly2to7_TS*) ;
- d'un groupe de 8 tissus ou plus (notés *poly8to47_TS*).

Afin d'observer si les profils d'expression étaient conservés pour les gènes présumés orthologues entre l'humain et la poule, les médianes d'expression des gènes (en TPM) issues des données RNAseq du GTEx V8 comptabilisant 53 tissus ont été utilisées [454].
La liste des gènes liés à un caractère mendélien connu a notamment été utilisée pour tester la concordance entre le tissu supposément impliqué et de l'expression du gène associé. Cette liste a été obtenue à partir du catalogue OMIA (*Online Mendelian Inheritance in Animals*) [127] et une réassignation manuelle a été faite pour certains gènes mis à jour dans l'assemblage GRCg7b.
Dans la même optique, la base de données miRNATissueAtlas2 a été exploitée pour quantifier l'expression des miARN chez l'humain [441] et nommant pour les miARN orthologues entre la poule et l'humain. Du fait de la difficulté à associer les miARN orthologues, c'est l'expression du précurseur du miARN qui a été utilisée.

Dans chaque tissu de chaque projet avec au moins huit individus par sexe, une analyse de l'expression différentielle (DE) entre sexes a été conduite avec le package edgeR [329] selon un modèle binomial négatif généralisé [455]. Les p-values ont été corrigées pour les tests multiples par la méthode de Benjamini-Hochberg [456] pour contrôler le taux de fausse découverte (FDR pour *False Discorvery Rate*), avec un seuil à 5%.

Les transcrits des PCG, lncRNA, miARN et snARN ont été classifiés par rapport à leur PCG ou lncRNA le plus proche, via la fonction FEELnc_classifier de FEELnc (v0.2.1) [122], avec une fenêtre maximum de 100 kb. La classification au niveau des modèles de gènes a été faite en combinant les résultats au niveau transcrit avec la fonction tpLevel2gnLevelClassification de FEELnc. Pour chaque paire lncRNA:PCG, lncRNA:lncRNA et PCG:PCG, la corrélation de Kendall (τ) entre les expressions à travers les tissus a été calculée et les p-values ont été corrigées pour

les tests multiples par la méthode de Benjamini-Hochberg [456]. Un FDR de 5%, correspondant à un $|\tau| \geq 0,55$ a ainsi été considéré pour qualifier des gènes comme co-exprimés.

Afin de valider certains modèles de gènes, des lncRNA d'intérêt et issus de différentes bases de données ont été testés expérimentalement par RT-PCR.

1.1.5. Valorisations associées

Ces travaux ont fait l'objet :

- d'un article en relecture par les paires : **Degalez F**, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F, Acloque H, Giuffra E, Pitel F, Lagarrigue S. (2023). Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues. bioRxiv. doi: 10.1101/2023.08.18.553750. **Cet article a été soumis à Scientific Reports. En attendant son traitement, il a été déposé sur bioRxiv. Il est reproduit ci-après ;**

- d'une communication orale (*) : **Degalez F**, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F, Giuffra E, Zerjal T, Pitel F, Lagarrigue S. (July 2023). A lncRNA gene-enriched atlas for GRCg7b chicken genome and its functional annotation across 47 tissues. Communication faite dans la session spécialisée "Avian Genetics and Genomics" au 39ème congrès de l' "International Society for Animal Genetics" (ISAG), Cape Town, South Africa ;

- d'un poster présenté à deux congrès :

    - (*) **Degalez F**, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F, Giuffra E, Zerjal T, Pitel F, Lagarrigue S. (July 2023). A lncRNA gene-enriched atlas for GRCg7b chicken genome and its functional annotation across 47 tissues. Communication faite dans la session spécialisée "Avian Genetics and Genomics " au 39ème congrès de l'International Society for Animal Genetics " (ISAG), Cape Town, South Africa. **Ce poster est reproduit ci-après à la suite de l'article ;**

    - **Degalez F**, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F, Giuffra E, Zerjal T, Pitel F, Lagarrigue S. (Sept 2023). A lncRNA gene-enriched atlas for GRCg7b chicken genome using Ensembl, RefSeq and two FAANG database. Communication faite dans la session spécialisée "EuroFAANG: genotype-to-phenotype research across Europe and beyond" pour le "74th European Federation of Animal Science Meeting" (EAAP), Lyon, France.

(*) Ces valorisations faites dans le cadre de l'ISAG 2023 ont fait l'objet d'une « *Travel Bursary Award* »

ARTICLE

Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues.

Fabien Degalez[1], Mathieu Charles[2], Sylvain Foissac[3], Haijuan Zhou[4], Dailu Guan[4], Lingzhao Fang[5], Christophe Klopp[2], Coralie Allain[1], Laetitia Lagoutte[1], Frédéric Lecerf[1], Hervé Acloque[6], Elisabetta Giuffra[6], Frédérique Pitel[3] and Sandrine Lagarrigue[1*]

[1]PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France.

[2]SIGENAE, INRAE, 31326 Castanet-Tolosan, France

[3]GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France

[4]University of California Davis, USA

[5]Aarhus University, Denmark

[6]Paris-Saclay University, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France;

Fabien Degalez: fabien.degalez@inrae.fr

Mathieu Charles: mathieu.charles@inrae.fr

Sylvain Foissac: sylvain.foissac@inrae.fr

Haijuan Zhou: hzhou@ucdavis.edu

Dailu Guan: dguan@ucdavis.edu

Lingzhao Fang: lingzhao.fang@qgg.au.dk

Christophe Klopp: christophe.klopp@inrae.fr

Coralie Allain: coralie.allain@institut-agro.fr

Laetitia Lagoutte: laetitia.lagoutte@inrae.fr

Frédéric Lecerf: frederic.lecerf@institut-agro.fr

Hervé Acloque: herve.acloque@inrae.fr

Elisabetta Giuffra: elisabetta.giuffra@inrae.fr

Frédérique Pitel: frederique.pitel@inrae.fr

Sandrine Lagarrigue (Corresponding author): sandrine.lagarrigue@institut-agro.fr

ABSTRACT

Gene atlases for livestock are steadily improving thanks to new genome assemblies and new expression data improving the gene annotation. However, gene content varies across databases due to differences in RNA sequencing data and bioinformatics pipelines, especially for long non-coding RNAs (lncRNAs) which have higher tissue and developmental specificity and are harder to consistently identify compared to protein coding genes (PCGs). As done previously in 2020 for chicken assemblies galgal5 and GRCg6a, we provide a new gene atlas, lncRNA-enriched, for the latest GRCg7b chicken assembly, integrating "NCBI RefSeq", "EMBL-EBI Ensembl/GENCODE" reference annotations and other resources such as FAANG and NONCODE. As a result, the number of PCGs increases from 18,022 (RefSeq) and 17,007 (Ensembl) to 24,102, and that of lncRNAs from 5,789 (RefSeq) and 11,944 (Ensembl) to 44,428. Using 1,400 public RNA-seq transcriptome representing 47 tissues, we provided expression evidence for 35,257 (79%) lncRNAs and 22,468 (93%) PCGs, supporting the relevance of this atlas. Further characterization including tissue-specificity, sex-differential expression and gene configurations are provided. We also identifiend conserved miRNA-hosting genes with human counterparts, suggesting common function. The annotated atlas is available at www.fragencode.org/lnchickenatlas.html.

Keywords: gene atlas, long non coding RNAs, chicken, genome annotation, tissue specificity, co-expression, *miRNA*

INTRODUCTION

Knowing the chromosomal gene content (*i.e.*, expressed regions) of an organism is crucial for most genetic studies including genetic responses of individuals or tissues to environmental variations, but also for identifying genes and genetic variants responsible for traits or diseases of interest. However, while protein coding genes (PCGs) are relatively well known, gene loci associated to long non-coding RNAs (lncRNAs) are more poorly described. LncRNAs, which have been widely described in the human genome in the early 2010s [1], are known to be gene expression regulators through various mechanisms, ranging from chromatin structure modification to transcription including RNA splicing regulation. They are also involved in RNA stability and translation [2,3] and therefore participate in various biological processes at the cellular and organism level [3–5]. Consequently, a comprehensive map of coding and non-coding transcribed regions is required to understand genotype to phenotype relationships. As an example, the human and mouse "EMBL-EBI Ensembl/GENCODE" (abbreviated as "Ensembl") genome annotations comprise 19,827 and 22,104 PCGs but 18,882 and 11,621 lncRNAs, respectively [6,7]. These known lncRNA counts is likely to increase as research [8,9]. For livestock species, lncRNAs are more and more integrated in reference genome annotations like "Ensembl" or "NCBI-RefSeq" (abbreviated "RefSeq") even if these catalogs are still very incomplete. We have previously shown discrepancies between these annotations in terms of transcript and gene models, strongly emphasizing variations for both lncRNAs and PCGs [10]: PCG models mainly differ at the transcript model level whereas lncRNA gene models differ both at the transcript and gene loci levels. Gene loci differ greatly between annotations, mainly due to specific features of lncRNAs (low expression, high tissue- and condition- specificity, …) and to the limited number of RNA-seq samples used to generate these catalogs. To facilitate accurate full-length transcript model reconstruction, annotation centers benefit from new technologies providing long-read RNA sequencing with an increase in accuracy and throughput, as well as a decrease in cost over time [11]. However, to properly detect lncRNAs, the high cost and so the low sequencing depths of these long-read technologies compared to short-read RNA-seq often require preliminary capture strategies to improve the concentration of low-abundance transcripts in cDNA libraries. This was successfully performed on human and mouse tissues by the GENCODE consortium [12]. Genome annotation databases are mainly supplied by short-read RNA-seq generated massively by the scientific community. In this context, to improve genome annotation

completeness, especially for lncRNAs, one strategy is to combine both most popular reference genome annotations – "RefSeq" and "Ensembl" – and other additional databases.

In this context, we provided in 2020 [13] a chicken atlas integrating gene models from "Ensembl", "RefSeq" and other databases. However, since 2020, the new GRCg7b chicken genome assembly with its associated reference genome annotations have been released, leading us to update and improve this gene atlas. Consequently, we included new databases such as FAANG multi-tissue resources and the NONCODE database, and provided an extensive functional annotation for the 24,102 PCGs and 44,428 lncRNAs using 1,400 RNA-seq samples from 47 tissues or cell types (available at www.fragencode.org/lnchickenatlas.html).We analysed their expression profile and provided a formatted table enabling easy extraction of the tissue(s) in which a gene of interest is the most expressed, notably to orient experimental studies. Furthermore, assuming that a gene expressed in a tissue or group of tissues plays a role related to the functions [14], we performed an in-depth analysis of lncRNA and PCG tissue-specific expression. We showed that lncRNAs are more tissue-specific than PCGs and illustrated the consistency between the expected and observed tissue specificity of genes involved in known Mendelian traits. We also provided a table of lncRNAs and PCGs hosting miRNA genes. We highlighted interesting cases, also conserved in human, in which both chicken and human lncRNAs expression profiles were similar and consistent with the miRNA function, suggesting a common function [15]. Finally, we classified lncRNAs based on their genomic configuration with respect to their closest PCG, defining lncRNA:PCG pairs. These pairs were then analysed in terms of co-expression across tissues since such a co-expression may be an indicator of a regulatory role of the lncRNA on the PCG [2,16–18], and therefore of their involvement in a common biological function, according to the "guilt-by-association" principle [19].

In summary, we provide a functional and genomic gene annotation table. Functional annotation includes various features such as the official short gene name, full gene name, identifier(s) and name(s) of human and mouse orthologous genes, expression profiles across 47 tissues and cell types, tissue specificity score, co-expression of lncRNA:PCG pairs, and other criteria. Genomic information provides the position of the genes and transcripts, the exon and intron numbers, the closest lncRNA or PCG, the overlap with a miRNA gene, and so forth. The extended gene model catalogue (*.gtf*) with coordinates on the GRGg7b genome assembly plus functional and genomic information (*.txt*) are available in this article

(Sup. Table 1), on the Fr-AgENCODE website (www.fragencode.org/lnchickenatlas.html) and on the dedicated interactive website (termed GEGA, gega.sigenae.org). Note that the files found on the website will be periodically updated with each novel significant chicken genome assembly version as already done for galgal5, GRCg6a and GRCg7b.

RESULTS

Overview of the different databases used to generate the chicken gene-enriched atlas.

Six databases containing lncRNAs and PCGs – for five of them – have been selected to create an enriched genome annotation. This set includes: *i)* "NCBI-RefSeq" (abbreviated in "RefSeq") and "EMBL-EBI Ensembl/GENCODE" (abbreviated in "Ensembl") databases, that are frequently updated and widely considered as references; *ii*) two databases from FAANG multi-tissue projects, namely the Fr-AgENCODE annotation ("FrAg") and the UC Davis annotation ("Davis"); *iii*) the INRAE annotation ("Inrae") previously used in Jehl et al., 2020 [13], for the gene enriched-atlas according to the GRCg6a assembly; *iv*) the "Noncode" database dedicated to non-coding RNAs. Comparison of the content of the gene models in the databases (Figure 1A) shows that PCGs overlap more with each other between databases than lncRNAs do. Thus, for PCGs, the "RefSeq"/"Ensembl"/"FrAg" dataset trio shows a high overlap rate around 95% globally, while consistency drops with the other annotations (75% for "Davis" and around 50% for "Inrae"). For lncRNAs, the overlap rate ranges from 50% for "RefSeq"/"FrAg" to 7% for "Ensembl"/"Davis". Note that despite their reference status, the overlap rate does not exceed 37% between both "RefSeq" and "Ensembl" reference databases. Consequently, as indicated by the low percentage of lncRNA overlapping, but also by the admittedly high but lower than 100% for PCGs, these resources appear complementary. As shown in Figure 1B-top (and Sup. Table 2), while PCG numbers are quite constant globally, with 18,022, 17,007, 14,078 and 18,341 for "RefSeq", "Ensembl, "FrAg" and "Davis" respectively, the number of lncRNAs is more variable, ranging from 5,789 at minimum for "RefSeq" to over 10,000 for the other databases, with, interestingly, a higher proportion of mono-exonic lncRNAs for "Inrae" and "Davis" (more than 65% against less than 24% for the other databases). This gene model variability between the six databases is also observed at the transcript level through the number of transcripts, which supports gene models (Figure 1B-bottom). Overall, the number of transcripts per gene is higher for PCGs than for lncRNAs and shows a greater variability. While the median number of transcripts is between 1 to 3 across the databases for PCGs, it does not exceed 1 for lncRNAs. As the number of transcripts supporting gene locus still low, regardless the PCG or lncRNA biotype, we chose to focus more on gene loci, level at which expression analyses are mostly performed, than on transcripts.

Based on these observations, we integrated the various annotations by sequentially adding gene loci from each database, keeping only gene loci that had no overlapping transcripts with

transcripts already present in the growing catalog (see Mat. & Meth. for more details). Since the conserved gene models in the enriched genome annotation – with their associated transcript models – are the ones that appear first during the successive additions of annotations, the gathering order is crucial. To better characterize the precision of transcripts models from each database, we computed the concordance between the annotated transcription start sites (TSS) and CAGE peaks from the FANTOM project (see Mat. & Meth.) (Figure 1C). The resulting support was higher for PCG promoters than for lncRNAs: the overlap rate between TSS and CAGE peaks varies between 60% for "RefSeq" and 40% for the other databases for PCGs (except for "Davis" which reaches 15%) whereas this overlap rate do not exceed 15% for lncRNAs. However, the rank of each database with respect to CAGE peaks is preserved, except for "Ensembl" that is lower for lncRNAs with only 5% of concordance.

Considering gene model quality characteristics (*i.e.*, number of gene loci and transcripts, biotypes, mono-exonicity), the concordance with CAGE peaks, and the popularity of each databases, the following order was chosen: 1-"RefSeq", 2-"Ensembl", 3-"FrAg", 4-"Davis", 5-"Inrae", 6-"Noncode". Consequently ad by construction, "RefSeq" gene models are fully included in the enriched genome annotation. Finally, this enriched gene atlas contains respectively 24,102 PCGs and 44,428 lncRNAs. Similarly, 991 miRNAs and a total of 78,323 gene models of various biotypes are annotated (Figure 1D & Sup. Table 3). This enriched gene atlas is available as a *.gtf* file on the Fr-AgENCODE website (www.fragencode.org/lnchickenatlas.html).

Interestingly, the PCG and lncRNA gene density per chromosome is correlated ($R = 0.62$, $p_{val} = 10^{-5}$) with a higher gene density in micro-chromosomes, which are better annotated since the GRCg7b update (Figure 1E). We observed 41 lncRNAs and 18 PCGs per Mb in macro-chromosomes (chr. 1-5) versus 66 for both in micro-chromosomes (chr. 11-39).

Gene expression across 47 chicken tissues.

In order to functionally characterize the 78,323 gene models, especially PCGs and lncRNAs, their expressions were quantified through 47 tissues (40 tissues *stricto sensu* and 7 cell types) coming from 36 datasets for a total of 1,400 individuals (see Mat. & Meth.), as presented in the Figure 2A and Sup. Table 4. This whole dataset is not exhaustive but tends to represent an important part of the physiology of the chicken by including tissues representing different specific systems such as the nervous (shades of grey), digestive (shades of green), respiratory

(shades of purple), sexual (shades of pink), circulatory (shades of brown), immune (shades of blue), or metabolic/energetic systems (shades of red).

A total of 63,513 (81%) genes are considered as expressed (Figure 2B-top), considering *inter alia*, a normalized expression threshold of 0.1 TPM and TMM (see Mat. & Meth.). This includes 22,468 (93%) PCGs and 35,257 (79%) lncRNAs. Interestingly, among the 6,238 genes with no defined biotype, identified as "other", 4,490 (72%) are also considered as expressed. The number of expressed genes per source (Figure 2B-bottom) averaged 75% but varied from 91% for “RefSeq” to 49% for “Noncode”, which is below the other databases due to older gene models and its addition as a final step in the sequential aggregation of gene models.

Regardless of the biotype, the PCAs performed on the expression data (Figure 2C and Sup. Figure 1), resulted in a tissue-dependent clustering across all datasets, validating the consistency of the expression data. Interestingly, lncRNAs clustered first the data according to the tissues with the most tissue-specific genes, *i.e.*, testis, brain and immunity (two sub-groups) related tissues. Moreover, considering all expressed genes, the 47 tissues are globally well classified across 14 classes with common biological functions (Figure 2D).

However, depending on the considered biotype and expression threshold, the number of expressed gene is variable: 88% (19,819/22,468) of PCGs have an expression ≥ 1 TPM in at least one tissue against 57% (20,252/35,257) of lncRNAs. In details, for a threshold of 0.1 TPM, the number of expressed PCGs varies from 9,887 (43.8%) in the caecal tonsils to 17,747 (78.6%) in utricle with an average of 14,837 (65.6%) PCGs expressed per tissue. Interestingly, the number of PCGs expressed in all tissues reached 7,435, *i.e.*, 75% of PCGs of the tissue with the lowest number of expressed PCGs and 33% of PCGs considered expressed in at least one tissue. For lncRNAs, a higher variability between tissues is observed. The number of expressed lncRNAs ranges from 1,189 (3.3%) for the caecal tonsil to 16,708 (46.5%) in testis with an average of 7,646 (21.3%) lncRNAs expressed per tissue. The number of lncRNAs expressed across all tissues reaches only 103, *i.e.*, 9% of lncRNAs for the tissue with the lowest number of expressed lncRNAs and 0.3% of lncRNAs considered expressed in at least one tissue. An expression threshold at 1 TPM lowers the average of expressed PCGs to 11,139 (FC = 1.3) and sharply drops the average of expressed lncRNAs to 1,972 (FC = 3.9) indicating that, as expected, lncRNAs are less expressed than PCGs within each tissue. All figures of PCGs and lncRNAs per tissue expressions are provided in Sup. Figure 2 and Sup. Table 5.

Tissue specific expression across 47 chicken tissues.

The tissue specificity, computed by the tau value (τ), seems to vary according to the expression threshold applied to consider a gene as expressed. For instance, considering a threshold of 0.1 and 1 TPM in at least one tissue, 86% (15,276/17,654) and 46% (18,417/40,071) of genes are tissue-specific (TS), respectively. According to this, we chose to work only with genes with an expression ≥ 1 TPM in at least one tissue (20,252 lncRNAs and 19,819 PCGs). PCGs and lncRNAs show different τ-values distributions, with lncRNAs globally more TS than PCGs, as already reported in Jehl et al., 2020, for chicken and dog [13]. Indeed, 23% (4,631) of PCGs have a $\tau \geq 0.9$ against 68% (13,786) for lncRNAs (Figure 3A). A local maximum around $\tau = 0.4$ is specifically observed for PCGs, suggesting more ubiquitously expressed genes.

Interestingly, genes that are considered as TS based on their tau value with an expression ≥ 1 TPM in at least one tissue, can still be expressed in several tissues, with highly variable expression profiles across tissues. For instance, by comparing for each TS gene the expression in the two tissues with the highest expression, resulting fold-change values can range from 1 to 10^5 TPM. To consider these cases, TS genes were split into three categories according to the expression pattern across the 47 tissues ("mono_TS", "poly2to7_TS", and "poly8to47_TS", see Mat. & Meth). Results showed that 3,378 (73%), 1,073 (23%) and 180 (4%) PCGs were specific to a unique tissue, a set of n tissues ($n \leq 7$) or without a specific group ($n > 7$), respectively. Same proportions were obtained for lncRNAs with 9,858 (72%), 3,225 (24%) and 703 (5%) genes, respectively (Figure 3B). More precisely, Figure 3C (top) indicates that the proportion of TS genes of each categories was very variable across tissues. As an example, 74% of the 4,905 genes which are TS for "testis" are mono-specific. In contrast, 0.8% of the 510 TS genes for the "duodenum" are mono-specific (all numbers per tissue are provided in Sup. Table 5). This variability in proportion is related to the other tissues present in the dataset and their common functions. Thus, tissues belonging to a common function tend to be express concomitantly for poly-specific genes. For example, tissues associated to the intestinal system as the duodenum, jejunum, ileum, cecum and colon, tend to express genes concomitantly as do the tissues associated to the brain system (Figure 3C-bottom). Thus, it should be noted that a gene can be considered as TS despite that no break in its expression pattern is observed. However, the opposite is also possible, a gene may be highly expressed in a tissue without being TS. For example, in the liver, one of the 5 most highly expressed PCG was not identified as TS as well as 5 of the top 15.

To illustrate the interest of gene expression tissue patterns, we examined expression profiles of causal genes associated to Mendelian traits. Out of the 54 Mendelian traits referenced by OMIA [20], 36 have strong hypotheses regarding the tissue in which the causal gene/variant was likely to affect, according to the trait's name or to the associated literature (Sup. Table. 6). Out of these 36 traits, 17 had a causal gene where one of the top two tissues with the highest expression was consistent with the tissue hypothesis. Some examples are shown in Figure 4A: *i)* GNB3, encoding a cone transducing subunit, causal gene of "Retinopathy globe enlarged" [21,22] with a retina-specific expression; *ii)* RBP, causal gene of "Riboflavin-binding protein deficiency" associated to embryonic death, with a magnum-specific expression which is consistent with the function of riboflavin-binding protein that transports the water-soluble vitamin from the oviduct into the egg white and also from serum into oocyte [23,24]; *iii)* KRT75L4, causal gene of "Frizzle, KRT75L4-related" responsible for a developmental defect of the feather [25,26] with a skin-specific expression. An intriguing case is the "LOC430486" gene *(iv)* responsible for chicken epilepsy [27,28] and encoding the synaptic vesicle glycoprotein 2A (SV2A) acting in the brain-related tissues. This gene was initially identified in 2011 in the galgal2 assembly (chr25:776,500-777,079 – 1 exon [29]) before being removed in subsequent releases because it was no longer predicted. It then reappeared in the galgal5 assembly both in "RefSeq" (LOC101748017) and in "Ensembl" (ENSGALG00000044909) but with a different gene structure and notably on a scaffold (KQ759566.1:4,207-4,692). Gene models were later harmonized between the two databases in the GRCg6a assembly (LOC101748017/ENSGALG00000050830) and the gene returned to its original position (chr25:1,854,812-1,880,902). It is also present in the GRCg7b assembly (LOC101748017/ENSGALG00010028753) for which, interestingly, the originally predicted sequence has a unique hit with 100% identity to the gene. Even if it is not tissue specific ($\tau = 0.81$), it is highly expressed in the cortex, brain, hypothalamus and cerebellum like its human ortholog (ENSG00000159164.9; $\tau = 0.58$; Figure 4B).

However, some traits deserve a more in-depth analysis, as illustrated by the blue eggshell. This trait for which the expected "causal" tissue should be uterus (the tissue responsible for eggshell formation) has for causal gene, SLCO1B3 [30] which is liver specific ($\tau = 0.95$) like its human ortholog (ENSG00000111700.12; $\tau = 0.98$), tissue where the associated protein transports a wide range of substrates including bile salts. The blue eggshell is due to a variant that leads to an ectopic expression of SLCO1B3 in uterus [31].

Differential expression between sexes.

We also provide a list of 4,206 differentially expressed genes (DEG) between sexes. These genes were identified in six tissues for which at least eight birds per sex were available from the same dataset: 2,475, 1,003, 768, 759, 659, and 233 DEGs were identified for liver, adipose tissue, bone marrow-derived macrophages, bursa of Fabricius, feathers, and the Harderian gland respectively (Sup. Table 7). These genes exhibited sex-biased expression in at least one of the six tissues, and correspond to 816 lncRNAs, 3,276 PCGs (*i.e.*, 8.3% and 19.9% of the total lncRNAs and PCGs expressed respectively) and 114 other gene biotypes. Of these, 3,384 (80.5%) genes are tissue-specific, *i.e.* sex-biased in only a single tissue, with similar percentages for lncRNAs (85.9%) and PCGs (79.5%). Most of these tissue-specific sex-biased PCG (75.7%) are expressed in more than three analysed tissues, this percentage is lower for lncRNAs (36.8%) (Figure 4C). The majority (691/822 genes, 84.1%) of genes showing sex-bias in two tissues or more has consistent fold-change directions between tissues. Of the 4,206 sex-biased genes, we observed an enrichment of Z-linked genes (821 genes, 19.5%) whereas only 5% of the total expressed genere Z-linked. They are characterized by a lower percentage of sex-biased expression in a single tissue (383 genes, 46.6%) compared to total DEG. As shown in Figure 4D, the incomplete sex chromosome dosage compensation known in chicken was observed with a median of log(fold-change "male/female") reaching 0.76. As for autosomal genes, the majority (419/438 genes, 95.7%) of Z-linked genes with sex biased expression in more than one tissue exhibit consistent effect directions across tissues.

LncRNAs host miRNA genes.

Using FEELnc, we classified the 991 chicken miRNAs into positional categories relatively to their closest lncRNA or PCG. We found that 244 (24.6%) and 717 (72.4%) miRNAs are hosted within an intron or an exon of 194 lncRNAs and 627 PCGs respectively. For lncRNAs, 43.8% (107) of miRNAs are within an intron against 51.6% (126) within an exon; for PCGs, 65.4% (469) are within an intron against 32.8% (235) within an exon. Note that 34 lncRNAs and 68 PCGs host more than one miRNA (six at most). Of the 194 lncRNAs, 77 (40%) come from the four resources excluding "RefSeq" and "Ensembl". Focusing on the 179 lncRNAs which are expressed (*i.e.*, TPM ≥ 0.1) in at least one tissue (hosting 228 miRNA), 133 (74.3%) have an expression ≥ 1 TPM, a significantly higher proportion compared to the expected proportion with total lncRNA (74.3% vs. 56.3%, $\chi 2$ = 1.6e-06); the same tendency was found for the 622 expressed PCGs associated to 712 miRNAs (98.2% vs. 87.8%, $\chi 2$= 2.2e-16). Out of the 179 lncRNAs, 110 (61.5%) are tissue-specific (same proportion as total

lncRNA) with 59, 19 and 13 lncRNAs specifically expressed in 1, 2 and 3 tissues, respectively. As expected, this tissue specific rate for lncRNAs is higher than that observed for PCGs for which only 61 genes (9.8%) are tissue specific. Except for MIR155HG, gene names of chicken lncRNAs hosting miRNA(s) are not standardized. We then observed tissue specific cases for miRNA hosted by lncRNA which are conserved in human with consistent tissue patterns between both species. For example, LOC124417505 (ENSGALG00010012701), hosting MIR122-1 within an exon, is identified as liver specific [29] like its human ortholog MIR122HG (Figure 5A). Similarly, LOC107052837 (ENSGALG00010019651), which hosts within an intron MIR217, is pancreas specific as its human ortholog MIR217HG. In addition, MIR217 is known to play a key role in pancreatic tumors [30]. Other tissue-specific lncRNAs which host miRNA(s) and newly modeled in this atlas also appear to be orthologous with known human lncRNAs hosting miRNA(s). For instance, NONGGAG008246, considered to be specific to the brain system, contains both gga-mir-219-a and gga-mir-219-b in an intron. Its presumed human ortholog, MIR219A2HG, also contains MIR219A and MIR219B [34, 35], all three specific to the brain system (Figure 5B).

Classification of the lncRNA with respect to the closest PCG and co-expression.

In order to detect biologically meaningful relationships between lncRNAs and PCGs based on the "guilt-by-association" principle [19], genes from both biotypes were classified according to their configuration with the closest PCG. Co-expressions between both genes constitutive of all lncRNA:PCG and PCG:PCG pairs were computed across the 47 tissues (Figure 5C). Out of the 35,257 lncRNAs and 22,468 PCGs considered as expressed, 33,907 (94,4%) and 20,656 (91,9%) are associated to a PCG within a 1 Mb window respectively (see Mat. & Meth). Out of them, 2,331 (6.9%) lncRNA:PCG pairs and 3,375 (16,4%) PCG:PCG pairs show a significant positive co-expression ($\rho \geq 0.55$; $pFDR \leq 0.05$). For all configurations, PCG:PCG pairs are more co-expressed than lncRNA:PCG pairs ($|\rho| = 0.16$ vs. 0.32). No negative and significant co-expressions were identified.

Thus, while coexpression can be used to generate hypotheses about the functionality of an lncRNA, the case of data from short-read sequencing must be considered with caution. Indeed, the length of the reads coupled with the low depth locally can sometimes lead to the erroneous modelling of new lncRNA genes (mono- or multi-exonic) upstream/5'UTR (untranslated transcribed region) or downstream/3'UTR of the PCG gene of the same strand, due to the inability to join adjacent genes. This phenomenon can lead to erroneous co-

expression and is expected to be more intense for downstream/3'UTR that are not well defined for PCG transcript models in our livestock species and can be much longer compared to the upstream/5'UTR. In line with this, we observed that lncRNAs in downstream/3'UTR of a PCG (noted "SS. down" – 12.6%) are more co□expressed with it compared to other intergenic configurations, especially lncRNAs in upstream/5'UTR (*i.e.*, "SS. up" – 5.2%) of a PCG. To illustrate these possible erroneous lncRNA model in downstream/3'UTR of a PCG, some lncRNA:PCG pairs in same strand, coming from different databases were tested by PCR for reliability. Three lncRNA:PCG pairs (LOC121113202/VSIG10L; NONGGAG001811/SARDH; FRAGALG000000006896/PA2G4) were identified in which the lncRNA was in the downstream/3'UTR of the PCG and can be considered as an extension of it. However, three other tested lncRNAs (DAVISGALG000044072/ADBR2 hosted, ENSGALG00010022678/PRPSAP2 in 5'UTR and ENSGALG00010016012/AMOT in 3'UTR) were found to be independent of the associated PCG (Sup. Figure 3).

Moreover, both lncRNAs and PCGs in "SS. up" and "divergent" configurations with another PCG show higher co-expression values than those in the "convergent" configuration. Excluding pairs in "SS. down" on focusing on intergenic pairs, we observed an enrichment in co-expressed genes ≤ 5kb compared to those ≥ 5kb for the "divergent" (11.6% vs. 3.0% for lncRNAs; 29.5% vs. 16.8% for PCGs) and "SS. up" configurations but not for the "convergent" one (1.8% vs 1.6% for lncRNAs; 10.5% vs. 9.5% for PCGs).

Overlap with the previous enriched annotation galgal5 and GRCg6a.

This work proposes a genome annotation (*.gtf*) and a gene annotation (*.tsv*) built on the GRCg7b assembly, that is considered as the new reference since April 2021 and July 2022 for "RefSeq" and "Ensembl" respectively [10]. This change in assembly and its coexistence with the previous GRCg6a and the alternate one GRCg7w, has led to a significant change in gene identifiers in some databases – particularly for "Ensembl" – which can complicate the transition and lead to uncertainties between studies performed on variable assemblies and annotations. For example, the SLC27A4 well-known protein coding gene is known as LOC417220 in "RefSeq" for galgal5, GRCg6a, GRCg7b and GRCg7w assemblies but in "Ensembl", the associated gene ID is ENSGALG00000004965 for galgal5 and GRCg6a, ENSGALG00010027394 for GRCg7b, and ENSGALG00015027711 for GRCg7w. To enhance the comparison between studies and different genome assemblies, we provide an equivalence table for *i*) the "Refseq" and "Ensembl" gene identifiers of GRCg7b for genes referenced in both databases, *ii*) the "Ensembl" gene identifiers of GRCg7b and GRCg7w,

iii) the gene identifiers from our previous annotation in galgal5 and GRCg6a to the one in GRCg7b (Sup. Table 8).

DISCUSSION

Our study proposes a solution for enriching the gene atlas of the two "RefSeq" and "Ensembl" chicken reference databases. This involves initially gathering these databases and then, supplementing them with four additional multi-tissue gene model resources, after determining a successive order of addition based on gene model quality criteria. While the use of a unique gene modeling pipeline including all raw sequencing data would be the best solution, our approach offers a good alternative. Indeed, *i)* it unifies the two most used genome annotations as the MANE (Matched Annotation from NCBI and EMBL-EBI) project which currently focused on the human [32] *ii)* it retains the identifiers of both "RefSeq" and "Ensembl" for common gene loci, *iii)* it is faster than a *de novo* annotation, and is adaptable to major changes in successive versions of the reference databases. Moreover, to facilitate the comparison between studies associated to different genome assemblies and genome annotations, we provided an identifier correspondence between galgal5 and GRCg6a to that of GRCg7b based on our previous gene-enriched model atlases anchored on the "Ensembl" genome annotation (v101 for GRCg6a; v94 for galgal5) [13,33]. This atlas increases the completeness of the chicken genome annotation, especially for lncRNAs, which are more difficult to identify than PCGs due to their low tissue- and condition-specific expression [8,34,35]. However, as the vast majority of current gene databases for livestock species are based on short-read data, transcript models are poorly described, regardless of gene biotype, even if this tendency is greater for lncRNAs than for PCGs [10]. As an example, across the six databases used in our study, the maximum median number of transcripts per lncRNA and PCG was one and three, respectively. These numbers are lower than those observed in human, with three and seven transcripts in average per lncRNA and PCG, respectively [10,34]. On the other hand, the overlap rate between transcript TSS and CAGE peaks, which are far from 100%, even for PCGs, underlines incomplete transcript modelling. These models will be clarified with long-read technologies, whose shortcoming today is the ability to obtain sequencing depths comparable to short-read technologies, thus limiting their massive use for studies focusing on gene expression [36]. Surprisingly, whereas the "Davis" database is the only one mainly based on long-read RNA-seq, we can note that this database has a poor overlap rate between TSS and CAGE for PCGs compared to other databases. Moreover, these lncRNAs, mainly mono-exonic, are generally located in the same strand of PCG introns, as for the "Inrae" ones. Indeed, 30.5% and 21.1% of the lncRNAs of "Davis" and "Inrae", respectively are in this case, a higher proportion compared to the other databases which oscillated between 4 and 7% (Sup.

Table 9). One interpretation could be that the low sequencing depth makes it difficult to build a full transcript model. Another limitation is that some gene loci can be erroneous, as illustrated in the manuscript, especially for lncRNAs that are on the same strand to a close PCG, and highly co-expressed. These lncRNAs could be in practice an untranslated transcribed region (UTR) of the PCG which are, as lncRNAs, challenging to model and need some complementary analyses [40, 41]. Therefore, PCR validation is required to verify the existence of such lncRNAs (*i.e.*, on the same strand to a close PCG) before further analyzing their functions using time-consuming molecular biology studies. Nevertheless, gathering genome annotations from multiple databases gives access to numerous new lncRNAs – precisely 44,428 lncRNAs including all the 5,789 and 11,944 loci from "RefSeq and "Ensembl" – since these datasets cover various tissues and conditions.

We then provide a gene annotation based on the expression across 47 tissues using 1400 samples from 36 datasets and found 81% of the gene models expressed in at least one tissue. As reported in the literature in cross-species analysis, lncRNAs are preferentially expressed in sexual tissues such as testis, potentially associated to a pervasive chromatin environment facilitating transcription of putatively non-functional elements enabling the emergence of new genes [39,40], and in a second time by tissue related to brain [1,41–43]. As expected, we found a higher tissue-specific proportion of lncRNAs compared to PCGs [1,13]. Expression profiles across tissues provide essential information for selecting relevant cell lines to study gene functions using different molecular biology methods [34]. It can also be a first indication of its function, especially for tissue-specific genes, as illustrated by the expression profile analysis of causal genes associated with Mendelian traits. However, it should be noted that tissue specificity is a relative measure, which depends on multiple factors including metric, threshold value or number of tissues. Among these factors, tissue specificity is particularly sensitive to the number and type of tissues. Adding another tissue can greatly vary gene tissue specificity values, especially when just few tissues are considered. Thus, the 40 tissues and 7 cell populations used in our study represent a strong resource. As an example, using a chicken dataset of 21 tissues, we showed in 2020 [13] a tissue specificity rate of 25% for lncRNAs vs. 10% for PCGs, against 68% and 23% observed respectively in this study. Tissue specificity also depends on the relationship between analyzed tissues, which explains why some genes are specific to several tissues, often sharing a similar functions.

We also provided a list of 4,206 genes with a sex-biased expression within six tissues corresponding to 19.8% of the total expressed PCGs, a lower percentage than reported by the

human GTEx consortium due to the higher number of analyzed tissues (37% of all genes with 44 tissues, [44]). Interestingly, 80% of sex biased genes are tissue-specific (sex DE observed in a single tissue), suggesting tissue-dependent regulation, even if this percentage is likely over-estimated in our study due to the low number of analyzed tissues (n = 6). This sex-biased tissue specificity does not reflect gene expression patterns across tissues since sex-biased genes tend to have ubiquitous expression across tissues, as previously reported by Oliva et al., 2020 [44]. Most of genes with sex biased expression in two or more tissues show consistent effect direction across tissues, especially for Z-linked genes, as previously reported [44]. Some genes reported in previous studies as differentially expressed between sexes in mammals have also been found in chicken: here some examples in liver with genes coding CYP3A4 related to drug metabolism [44,45], von Willebrand factor C and EGF domains (VWCE alias urg11) predicted to enable calcium ion binding activity [44,46], polycystin 2 (PKD2), a membrane protein involved in a calcium-permeant cation channel [46] or calcitonin-related polypeptide alpha (CALCA) [47].

Our findings indicate that most 991 chicken miRNAs are located within a gene, with 75% of them within a PCG and 25% within a lncRNA. These results are in line with those of Liu et al., 2018, who demonstrated that a large fraction of miRNAs in miRBase v21 (1325 out of 1881) are also hosted in a gene [48] and with those of Dhir et al., 2015, who reported, in human, a small fraction of miRNA (17.5%) hosted by a lncRNA [15]. Among the hosted chicken miRNAs, we observed that nearly all of them are embedded in an intron or an exon of its hosting gene. The location of miRNAs according to the nearest gene is an important factor to consider to investigate the transcriptional regulation of primary miRNAs, which is not yet fully understood. Previous studies in human have reported that more than half of miRNAs reside in PCG introns (no study focusing specifically on lncRNAs) and are thought to be co-expressed with their host genes, deriving from common primary transcripts [49–52]. This assumption needs to be moderated since Ozsolak et al., 2008, [53] reported that a significant fraction of intragenic miRNAs were independently initiated from the PCG transcripts. Additional data would be needed to test the co-expression of miRNA and its host gene. In the absence of the aforementioned data, miRNAs that exhibited conserved genomic localization with their host lncRNA and with a similar expression profiles in both human and chicken were analyzed. We then demonstrated that the expression profile of the host lncRNA matched that of the human, providing strong evidence of co-regulation. Notably, three cases of interest were highlighted, including MIR122-1, which is hosted by

LOC124417505/ENSGALG00010012701, MIR217 hosted by LOC107052837/ ENSGALG00010019651, and MIR219A and MIR219B hosted by NONGGAG008246, all corresponding to miRNAs nested within an intron or an exon. Gene names of chicken lncRNAs hosting miRNA(s) are not standardized and should be called MIRxxxHG as MIR155HG, the only lncRNA correctly named, following our work published in 2020 which provided a first functional annotation table of chicken genes related to the chicken genome assemblies, galgal5 and GRCg6a and which identified it as the INRAGALG00000001802 lncRNA [13].

Analyses of lncRNA:PCG configurations shows that lncRNAs tend to be more genic rather than intergenic. Although, while this observation may vary according to different sources [1,54] it can be explained by *i)* the use of unoriented RNA-seq data for the oldest publications, *ii)* the consideration of only multi-exonic transcript models by the bioinformatics pipelines to avoid potential false positives corresponding to poorly covered transcripts and, *iii)* the drop of short-read RNA-seq cost allowing now to sequence in greater depth and to better consider low expressed transcripts. In our study, we observed an over-evaluation of intragenic lncRNAs, which may be explained by the use of a long-read sequencing database, limited in depth. Focusing on intergenic genes and as shown in the literature [1,54], an enrichment in "same-strand" is observed. LncRNAs involved in such configurations should be considered with caution, since, as illustrated in the manuscript, some of them are part of a not well-modeled PCGs. Indeed, a lot of PCG isoforms are still poorly annotated, especially for non-model species. For example, as shown by Lagarrigue et al., 2021 [34], for a stable number of gene models, the number of PCG transcripts oscillates between 28,000 and 50,000 for farm species while it exceeds 100,000 for mouse and 150,000 for human. A very high co-expression value across tissues (or intra tissue according to the study) and a low distance between gene models can be considered as a distrust indicator. As an example, Muret et al., 2019 showed with a PCR validation that the FLRL7 lncRNA in "same-strand down" of FADS2 in the mouse constituted in reality a single gene model [5]. However, if some lncRNA:PCG pairs in "same-strand" must be considered with precaution, a considerable part of the constitutive lncRNAs seems to exist independently. Consequently, as well as for the "divergent" or genic lncRNA:PCG co-expressed pairs, it is possible to propose hypotheses concerning the lncRNA function applying the "guilt-by-association" principle [19]. Indeed, a significant expression correlation and a short distance between two gene models can supposed a common regulation or even an implication of the lncRNA in the regulation of the

PCG [55–59]. The co-expression of lncRNA:PCG pairs in "divergent" configuration could be related to a bidirectional-promoters which could activate the expression of the PCG through an alteration of the promoter regions by the lncRNA (named pancRNA for promoter-associated non-coding RNA) [55,60,61]. For example, Hamazaki et al., 2017 showed that the lncRNA pancIl17d, in "divergent" configuration with the PCG Il17d is crucial for pre-implantation development of mouse through an upregulation [62]. This pancRNA expression leads to a DNA demethylation and an upregulation to its associated PCG. Interestingly, across all the lncRNA:PCG and PCG:PCG configurations, no significant negative correlation was identified. Indeed, as observed in other species such as human [1], dog [63], and even in plants [35], only a tiny fraction of lncRNA:PCG pairs showed a significant negative co-expression. Even if some cases of silencing are well-known, this suggest that lncRNAs tend to act as positive regulators or cofactors improving the expression of near genes through various mechanisms [3]. Finally, considering all configurations, lncRNA:PCG pairs have lower co-expressed pairs across tissues compared to PCG:PCG. This observation highlights the tissue (and condition) specificity of lncRNAs compared to the ubiquity of PCGs [1,64,65]. Thus, in order to establish robust hypotheses about the association of function between a lncRNA and a nearby PCG, it is essential to consider the co-expression within the tissue(s) of interest and for a unique condition. The combined use of the configuration of lncRNA:PCG pairs and their co-expression can help to orient the hypotheses and the biological experiments to set up in order to better understand the regulatory functions of lncRNAs.

In conclusion, if your research field is focused on gene expression analysis in chicken and you use this enriched atlas, 24,102 PCGs and 44,428 lncRNAs containing all gene loci from "RefSeq and "Ensembl" instead of only 18,022 and 17,007 PCGs and 5,789 and 11,944 lncRNAs for "RefSeq and "Ensembl" respectively. Among them, note that 19,819 PCGs and 20,252 lncRNAs have an expression ≥ 1 TPM in at least one tissue, ensuring an easy handling for further investigation by molecular biology methods to gain insight into their function. For all these genes, we also provide a table containing different genomic and functional information/feature (Sup. Table. 1) soon available through a web interface. The atlas and related information will be valuable for researchers working on gene expression (PCGs and/or lncRNAs), such as those interested in unraveling the molecular mechanisms linking non-coding variants and relevant phenotypes.

METHODS

Reference assembly

The genome annotation was constructed according to the bGalGal1.mat.broiler.GRCg7b (GCF_016699485.2) assembly of the chicken (Gallus Gallus) genome [66].

Gene-enriched atlas construction

Origin of the six genome annotations. Gene models used to build the enriched genome annotation come from 6 genome annotations, all based on multi-tissue resources: *i)* both reference genome annotations according to the GRCg7b assembly: "RefSeq" v106 [67] and "Ensembl" v107 [68], this latter has integrated the GENESWitCH project data; *ii)* both gene model datasets from FAANG pilot projects [69,70] according to the GRCg6a assembly (GCF_000002315.5): the FR-AgENCODE project [71] involving 11 tissues represented by 2 males and 2 females by tissue and the FarmENCODE project including 15 tissues with 1 male and 1 female; *iii)* and two other datasets including gene models from the previous atlas as presented in Jehl et al., 2020 [13] produced according to the galgal5 assembly (GCF_000002315.4) and NONCODE v6.0 [72] including only non-coding gene models from the literature and other public databases according to the galgal4 assembly (GCF_000002315.2). Contrary to all projects which used short-read sequencing, FarmENCODE includes samples sequenced with Oxford Nanopore long read Technology as presented in Guan et al., 2022 [73]. For genome annotations produced on a previous assembly, a remapping to GRCg7b was performed using the NCBI genome remapping service [74].

Prioritization criteria. CAGE data used to prioritize the different gene models come from the FANTOM5 project [75]. Peaks coordinates considered as robust according to the project were converted from galgal5 (GCF_000002315.4) to GRCg7b using the NCBI genome remapping service [74]. The transcript is then considered to be well modelled in 5' if its TSS overlap a peak within +/- 30bp. All genome annotations previously presented were added sequentially considering gene model quality characteristics, the concordance with CAGE peaks, and the popularity of each databases as presented in the "Results" part, namely: 1-"RefSeq"; 2-"Ensembl"; 3-"FrAg"; 4-"Davis"; 5-"Inrae"; 6-"Noncode" (Figure 6A).

Rules of aggregation. Two gene models were considered overlapping if at least one of their transcripts had at least one of their exons with a base pair (1 bp) in common and on the same strand (Figure 6B). Overlapping detection was performed using the "intersect" function (parameters -wo -s) of the BEDTool v2.25.0 toolset [76]. To improve the successive addition of

the different gene models, a decomposition by biotype class was used (See Sup. Table 10). This approach limited the overlap of similar gene patterns, but with different or unassigned biotypes, and was more sensitive to genes hosting other genes, such as miRNA-hosting PCGs for example.

Biological sample used for gene expression

36 datasets including a total of 1400 samples were used to represent the 47 tissues composing the atlas. As these datasets are publicly available (on SRA and/or ENA), the project numbers and the number of samples are available in the Sup. Table 4.

The 47 tissues and their respective four letter abbreviations are: adipose tissue (adip), blood (blod), bone marrow derived macrophages (bmdm), brain (brai), bursa of Fabricius (burs), caecal tonsil (cctl), cecum (cecm), chorioallantoic membrane of an embryo (chor), colon (coln), cerebellum (crbl), cortex (crtx), dendritic cell (denC), duodenum (duod), embryon (ember), feather (feat), gizzard (gizz), Harderian gland (hard), heart (hert), hypothalamus (hypt), ileum (ileu), isthmus (isth), jejunum (jeju), kidney (kdny), liver (livr), lung (lung), lymphocyte B (lymB), lymphocyte T CD4 and CD8 (lymT), magnum (magn), monocyte, (mono), breast muscle (mscB), IEL-NK celles (nkil), optic lobe (optc), ovary (ovry), pancreas (pcrs), pineal gland (pine), pituitary (pitu), proventriculus (pvtc), retina (rtin), skin (skin), spleen (spln), testicule (test), thrombocyte (thro), thymus (thym), thyroid gland (thyr), trachea (trch), uterus (uter) and utricle (utri). Color codes associated to each tissue are available in Sup. Table 11.

Gene expression quantification and expression criteria

FASTQ files were mapped on the GRCg7b reference genome (GCF_016699485.2) and expression quantification according to the enriched *.gtf* annotation file was performed by projects and using the "rnaseq" v3.8.1 pipeline (--aligner star-rsem) from nf-core [77,78] providing raw counts and TPM normalized counts. For each tissue in each project, a median of TPM normalized expressions across samples was calculated. For tissues present in several projects, the median was calculated using the TPM medians previously calculated in each project.

A gene was considered as expressed if its median expression (see previous §) was ≥ 0.1 TPM in at least one tissue and if at least 50% of samples of a tissue for a given project have a reads number ≥ 6 and the normalized TPM and TMM expression ≥ 0.1. TMM normalized expression was obtained from the raw counts by the trimmed mean of M-values (TMM)

scaling factor method [79] using the R package edgeR (v3.32.1) [80] with the "calcNormFactors" function (to scale the raw library sizes) and "rpkm" function (to scale the gene model size). Finally, genes were classified into three expression categories: genes with expression *i*) < 0.1 TPM in all tissues *ii*) ∈ [0.1, 1[TPM in at least one tissue *iii*) ≥ 1 TPM in at least one tissue.

PCA and clustering

PCA was performed with the "PCA" function (scale.unit = T) of the FactoMineR (v2.7) [81] package and considering the $\log_2$(TPM+1) expression of the expressed genes. The dendrogram was based on the distance matrix computed with the (1-Pearson correlation) of the $\log_2$(TPM+1) expression of the expressed genes and the hierarchical cluster analysis was done using the "ward.D" agglomeration method and the "hclust" function.

Tissue-specificity analysis

Tissue-specificity was assessed with the $\log_{10}$ median expression of tissues. The tau (τ) metric was used [82], providing a score between 0 (gene expressed at the same level in all tissues) and 1 (gene expressed in exactly one tissue). A gene was considered as tissue specific for a $\tau \geq 0.90$ and in some analyses (related to Figure 3) with a filter on the expression (≥ 1 TPM in at least one tissue). Genes considered as tissue-specific ($\tau \geq 0.90$) were split into three categories based on the expression profile and whether or not a gap – define as a difference in expression by a factor of 2, *i.e.*, FC ≥ 2 – was observed between tissues expressions when they were ordered in descending order. The three categories of tissue specific expression were defined as follows: genes specifically expressed in *i)* a unique tissue (mono_TS), *ii)* a group of 2 to 7 tissues (included) (poly2to7_TS) or *iii)* a group of 8 or more tissues (poly8to47_TS).

GTEx data analysis

The median gene-level TPM for 53 tissues from RNA-seq data of GTEx Analysis V8 was used (https://gtexportal.org/home). The list of the 53 tissues, their abbreviations and color codes used are available in Sup. Table 12.

OMIA gene lists

Genes related to a known Mendelian trait or disorder were obtained from the OMIA (Online Mendelian Inheritance in Animals) catalog [20]. A manual reassignment was performed for C1H12ORF23, GC1, KIAA0586, LOC430486 genes that were updated in the GRCg7b assembly (Sup. Table 6).

Differential gene expression between sexes

First, the genes "expressed" in each tissue of each project for which at least 8 birds per sex were available were identified. The tissues and projects concerned were "hard – PRJNA484002", "burs – PRJEB23810", "bmdm – PRJEB34093", "bmdm – PRJEB22373" and 'livr, blod, adip – PRJEB44038". A gene was considered as expressed if the normalized TPM and TMM expressions were ≥ 0.1 and if the read counts was ≥ 6 in at least 80% of the samples of one sex. Then, the differential expression (DE) analysis using the raw counts of the expressed genes previously selected was performed using the R package edgeR (v3.32.1) [80] based on a generalized negative binomial model for model fitting. The "edgeR-Robust" method was used to account for potential outliers when estimating per gene dispersion parameters [83]. P-values were corrected for multiple testing using the Benjamini-Hochberg approach [84] to control the false discovery rate (FDR), and genes were identified as significantly differentially expressed if pFDR < 0.05. For the "bmdm" tissue where two projects were available, the DEG union was considered. List of DEG per tissues is available in Sup. Table 7.

miRNA expression in human

The "miRNATissueAtlas2" database was exploited to quantify the expression of miRNA for human [51]. Because of the difficulty in associating the orthologous miRNAs between the chicken and the human, the expression of the miRNA precursor was used.

Classification according to the closest feature

PCG, lncRNA, miRNA and snRNA transcripts were classified relatively to their closest PCG and lncRNA transcript using the "FEELnc_classifier" function of FEELnc v.0.2.1 with a maximum window of 100 kb (default setting) [85]. The classification for gene models was performed by combining the transcript results and the "tpLevel2gnLevelClassfication" function from FEELnc.

Co-expression analysis

For each lncRNA:PCG, lncRNA:lncRNA and PCG:PCG pairs, the Kendall correlation (τ) between the expression values across tissues was computed. Genes were considered as co-expressed for a $|\tau| \geq 0.55$ after that p-values were corrected for multiple testing using the Benjamini-Hochberg method [84] and applying a false discovery rate of 0.05.

Biological validation by RT-PCR

Reverse transcription (RT) was carried out using the high- capacity cDNA archive kit (Applied Biosystems, Foster City, CA) according to the manufacturer's protocol. Briefly, reaction mixture containing 2 µL of 10× RT buffer, 0,8 µL of 25X dNTPs, 2 µL of 10X random primers, 1 µL of MultiScribe Reverse Transcriptase (50 U/ µL), and total RNA (1 µg) was incubated for 10 min at 25 °C followed by 2 h at 37 °C and 5 min at 85 °C. RT reaction was diluted to 1/5 and further used for PCR. 5 µl of cDNA and 5µL of gDNA were mixed separately with 8µL of 5X Green or Colorless GoTaq Flexi Buffer, 3,2mL of MgCl2 25mM, 0,8µL of dNTPs 10mM, 15,8µL H20, 0,2µL of GoTaqG2 Hot Start Polymerase (5u/µl) and 500nM of specific reverse and forward primers. Reaction mixtures were incubated in an T100 thermal cycler (Bio-Rad, Marne la Coquette, France) programmed to conduct one cycle (95 °C for 3 min), 40 cycles (95 °C for 30 s, 61,5°C to 64°C for 30 s and 72 °C for 1 min to 3 min, depending on primers used) and a last cycle (72°C for 5 min). PCR products were mixed with loading dye and was run at 100 V for 35 min on 1.5% agarose gel. Primers sequences and the corresponding annealing temperature are provided in Sup. Table 13.

REFRENCES

1. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
2. Gil, N. & Ulitsky, I. Regulation of gene expression by cis-acting long non-coding RNAs. *Nat. Rev. Genet.* **21**, 102–117 (2020).
3. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
4. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641 (2009).
5. Muret, K. *et al.* Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. *BMC Genomics* **20**, 882 (2019).
6. EMBL-EBI Ensembl/GENCODE. GRCh38.p13 - Genome - Annotation - Ensembl v109. https://www.ensembl.org/Homo_sapiens/Info/Annotation (2023).
7. EMBL-EBI Ensembl/GENCODE. GRCm39 - Genome - Annotation - Ensembl v109. https://www.ensembl.org/Mus_musculus/Info/Annotation (2023).
8. Jiang, S. *et al.* An expanded landscape of human long noncoding RNA. *Nucleic Acids Res.* **47**, 7842–7856 (2019).
9. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* **19**, 535–548 (2018).
10. Smith, J. *et al.* Fourth Report on Chicken Genes and Chromosomes 2022. *Cytogenet. Genome Res.* 1 (2023) doi:10.1159/000529376.
11. Marx, V. Method of the year: long-read sequencing. *Nat. Methods* **20**, 6–11 (2023).
12. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* **49**, 1731–1740 (2017).
13. Jehl, F. *et al.* An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci. Rep.* **10**, 20457 (2020).
14. Odom, D. T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).

15. Dhir, A., Dhir, S., Proudfoot, N. J. & Jopling, C. L. Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. *Nat. Struct. Mol. Biol.* **22**, 319–327 (2015).
16. Luo, S. *et al.* Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell* **18**, 637–652 (2016).
17. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
18. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
19. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
20. Sydney School of Veterinary Science, University of Sydney. Online Mendelian Inheritance in Animals - OMIA. https://www.omia.org/ (2023).
21. Sydney School of Veterinary Science, University of Sydney. Retinopathy globe enlarged in Gallus gallus - OMIA. https://www.omia.org/OMIA001368/9031/ (2011).
22. Tummala, H. *et al.* Mutation in the Guanine Nucleotide–Binding Protein β-3 Causes Retinal Degeneration and Embryonic Mortality in Chickens. *Invest. Ophthalmol. Vis. Sci.* **47**, 4714–4718 (2006).
23. MacLachlan, I., Nimpf, J., White, H. B. & Schneider, W. J. Riboflavinuria in the rd chicken. 5′-splice site mutation in the gene for riboflavin-binding protein. *J. Biol. Chem.* **268**, 23222–23226 (1993).
24. Sydney School of Veterinary Science, University of Sydney. Riboflavin-binding protein deficiency in Gallus gallus - OMIA. https://www.omia.org/OMIA000876/9031/ (2022).
25. Dong, J. *et al.* A novel deletion in KRT75L4 mediates the frizzle trait in a Chinese indigenous chicken. *Genet. Sel. Evol. GSE* **50**, 68 (2018).
26. Sydney School of Veterinary Science, University of Sydney. Frizzle, KRT75L4-related in Gallus gallus - OMIA. https://www.omia.org/OMIA002486/9031/ (2021).
27. Douaud, M. *et al.* Epilepsy caused by an abnormal alternative splicing with dosage effect of the SV2A gene in a chicken model. *PloS One* **6**, e26932 (2011).
28. Sydney School of Veterinary Science, University of Sydney. Epilepsy in Gallus gallus - OMIA. https://www.omia.org/OMIA000344/9031/ (2011).

29. LOC430486 similar to Ca2+ regulator SV2A [Gallus gallus (chicken)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/430486.
30. Sydney School of Veterinary Science, University of Sydney. Blue eggshell in Gallus gallus - OMIA. https://www.omia.org/OMIA000142/9031/ (2022).
31. Wang, Z. *et al.* An EAV-HP Insertion in 5 Flanking Region of SLCO1B3 Causes Blue Eggshell in the Chicken. *PLOS Genet.* **9**, e1003183 (2013).
32. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).
33. FR-AgENCODE. FR-AgENCODE - functional annotation of livestock genomes. https://www.fragencode.org/ (2023).
34. Lagarrigue, S., Lorthiois, M., Degalez, F., Gilot, D. & Derrien, T. LncRNAs in domesticated animals: from dog to livestock species. *Mamm. Genome* **33**, 248–270 (2022).
35. Xu, Q. *et al.* Systematic comparison of lncRNAs with protein coding mRNAs in population expression and their response to environmental change. *BMC Plant Biol.* **17**, 42 (2017).
36. Soneson, C. *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* **10**, 3359 (2019).
37. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
38. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15776–15781 (2003).
39. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).
40. Soumillon, M. *et al.* Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Rep.* **3**, 2179–2190 (2013).
41. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510–514 (2019).
42. Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* **24**, 616–628 (2014).

43. Hezroni, H. *et al.* A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol.* **18**, 162 (2017).
44. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, eaba3066 (2020).
45. Rinn, J. L. & Snyder, M. Sexual dimorphism in mammalian gene expression. *Trends Genet. TIG* **21**, 298–305 (2005).
46. García-Calzón, S., Perfilyev, A., de Mello, V. D., Pihlajamäki, J. & Ling, C. Sex Differences in the Methylome and Transcriptome of the Human Liver and Circulating HDL-Cholesterol Levels. *J. Clin. Endocrinol. Metab.* **103**, 4395–4408 (2018).
47. Gershoni, M. & Pietrokovski, S. The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol.* **15**, 7 (2017).
48. Liu, B., Shyr, Y., Cai, J. & Liu, Q. Interplay between miRNAs and host genes and their role in cancer. *Brief. Funct. Genomics* **18**, 255–266 (2018).
49. Baskerville, S. & Bartel, D. P. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA N. Y. N* **11**, 241–247 (2005).
50. DOHI, O. *et al.* Epigenetic silencing of miR-335 and its host gene MEST in hepatocellular carcinoma. *Int. J. Oncol.* **42**, 411–418 (2012).
51. Cai, Y., Yu, X., Hu, S. & Yu, J. A brief review on the mechanisms of miRNA regulation. *Genomics Proteomics Bioinformatics* **7**, 147–154 (2009).
52. Kim, Y.-K. & Kim, V. N. Processing of intronic microRNAs. *EMBO J.* **26**, 775–783 (2007).
53. Ozsolak, F. *et al.* Chromatin structure analyses identify miRNA promoters. *Genes Dev.* **22**, 3172–3183 (2008).
54. Kern, C. *et al.* Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genomics* **19**, 684 (2018).
55. Wei, W., Pelechano, V., Järvelin, A. I. & Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet. TIG* **27**, 267–276 (2011).
56. Gibbons, H. R. *et al.* Divergent lncRNA GATA3-AS1 Regulates GATA3 Transcription in T-Helper 2 Cells. *Front. Immunol.* **9**, 2512 (2018).

57. Canzio, D. *et al.* Antisense lncRNA Transcription Mediates DNA Demethylation to Drive Stochastic Protocadherin α Promoter Choice. *Cell* **177**, 639-653.e15 (2019).
58. Rom, A. *et al.* Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nat. Commun.* **10**, 5092 (2019).
59. George, M. R. *et al.* Minimal in vivo requirements for developmentally regulated cardiac long intergenic non-coding RNAs. *Dev. Camb. Engl.* **146**, dev185314 (2019).
60. Uesaka, M., Agata, K., Oishi, T., Nakashima, K. & Imamura, T. Evolutionary acquisition of promoter-associated non-coding RNA (pancRNA) repertoires diversifies species-dependent gene activation mechanisms in mammals. *BMC Genomics* **18**, 285 (2017).
61. Uesaka, M. *et al.* Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. *BMC Genomics* **15**, 35 (2014).
62. Hamazaki, N., Uesaka, M., Nakashima, K., Agata, K. & Imamura, T. Gene activation-associated long noncoding RNAs function in mouse preimplantation development. *Dev. Camb. Engl.* **142**, 910–920 (2015).
63. Le Béguec, C. *et al.* Characterisation and functional predictions of canine long non-coding RNAs. *Sci. Rep.* **8**, (2018).
64. Jiang, C. *et al.* Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs. *Oncotarget* **7**, 7120–7133 (2016).
65. de Goede, O. M. *et al.* Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **184**, 2633-2648.e19 (2021).
66. NCBI-RefSeq. bGalGal1.mat.broiler.GRCg7b - Genome - Assembly - NCBI. https://www.ncbi.nlm.nih.gov/assembly/GCF_016699485.2/ (2021).
67. NCBI-RefSeq. bGalGal1.mat.broiler.GRCg7b - Genome - Annotation - NCBI v106. https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/699/485/GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b/ (2022).
68. EMBL-EBI Ensembl/GENCODE. bGalGal1.mat.broiler.GRCg7b - Genome - Annotation - Ensembl v107. https://ftp.ensembl.org/pub/release-107/gtf/gallus_gallus/ (2022).
69. Tixier-Boichard, M. *et al.* Tissue Resources for the Functional Annotation of Animal Genomes. *Front. Genet.* **12**, 666265 (2021).

70. Andersson, L. *et al.* Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **16**, 57 (2015).
71. Foissac, S. *et al.* Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* **17**, 108 (2019).
72. Zhao, L. *et al.* NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res.* **49**, D165–D171 (2021).
73. Guan, D. *et al.* Prediction of transcript isoforms in 19 chicken tissues by Oxford Nanopore long-read sequencing. *Front. Genet.* **13**, (2022).
74. Coordinate remapping service: NCBI. https://www.ncbi.nlm.nih.gov/genome/tools/remap.
75. Lizio, M. *et al.* Systematic analysis of transcription start sites in avian development. *PLoS Biol.* **15**, e2002887 (2017).
76. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
77. Patel, H. *et al.* nf-core/rnaseq: nf-core/rnaseq v3.8.1 - Plastered Magnesium Mongoose. (2022) doi:10.5281/zenodo.6587789.
78. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).
79. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
80. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
81. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
82. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinforma. Oxf. Engl.* **21**, 650–659 (2005).
83. Zhou, X., Lindsay, H. & Robinson, M. D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* **42**, e91 (2014).
84. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

85. Wucher, V. *et al.* FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **45**, e57 (2017).

AKNOWLEDGEMENTS

We would like to thank Sophie Rehault, who trusted us by sharing data that had not yet been made public at the time of the analysis. We are also grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi: 10.15454/1.5572369328961167E12) for providing help and/or computing and/or storage resources.

This project is funded by the European Union's Horizon 2020 research and innovation program under grant agreement N°101000236 (GEroNIMO) and by ANR CE20 under 'EFFICACE' program. FD is a Ph.D. student supported by the Brittany region (France) and the INRAE (Animal Genetics Division). These funding bodies had no role in the design of the study, in the collection, analysis, and interpretation of data, or in writing the manuscript.

AUTHOR CONTRIBUTIONS

FD and SL conceived and coordinated the study. FD, MC and SL performed bioinformatics processing of the RNA-seq data. SL acquired funding for this research. FD and SL carried out the whole bioinformatics analysis. CA and LL carried out the PCR analysis. FL was responsible for the computational infrastructure. FD and SL drafted the manuscript and figures. SF, HZ, DG, LF, CK, CA, LL, FL, HA, EG and FP helped to improve the manuscript. All authors reviewed and approved the final version.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

RNAseq data are publicly available at https://www.ebi.ac.uk/ena/browser/home and the corresponding project accession number are provided in the Sup. Table 4.

Genome annotation files are publicly accessible as referenced in the "Methods" section.

Data generated during this study are included in this published article (and its Supplementary Information files), on the https://www.fragencode.org/lnchickenatlas.html website and interactively using the https://gega.sigenae.org/ tool.

LEGENDS FIGURES AND TABLES

MAIN FIGURES

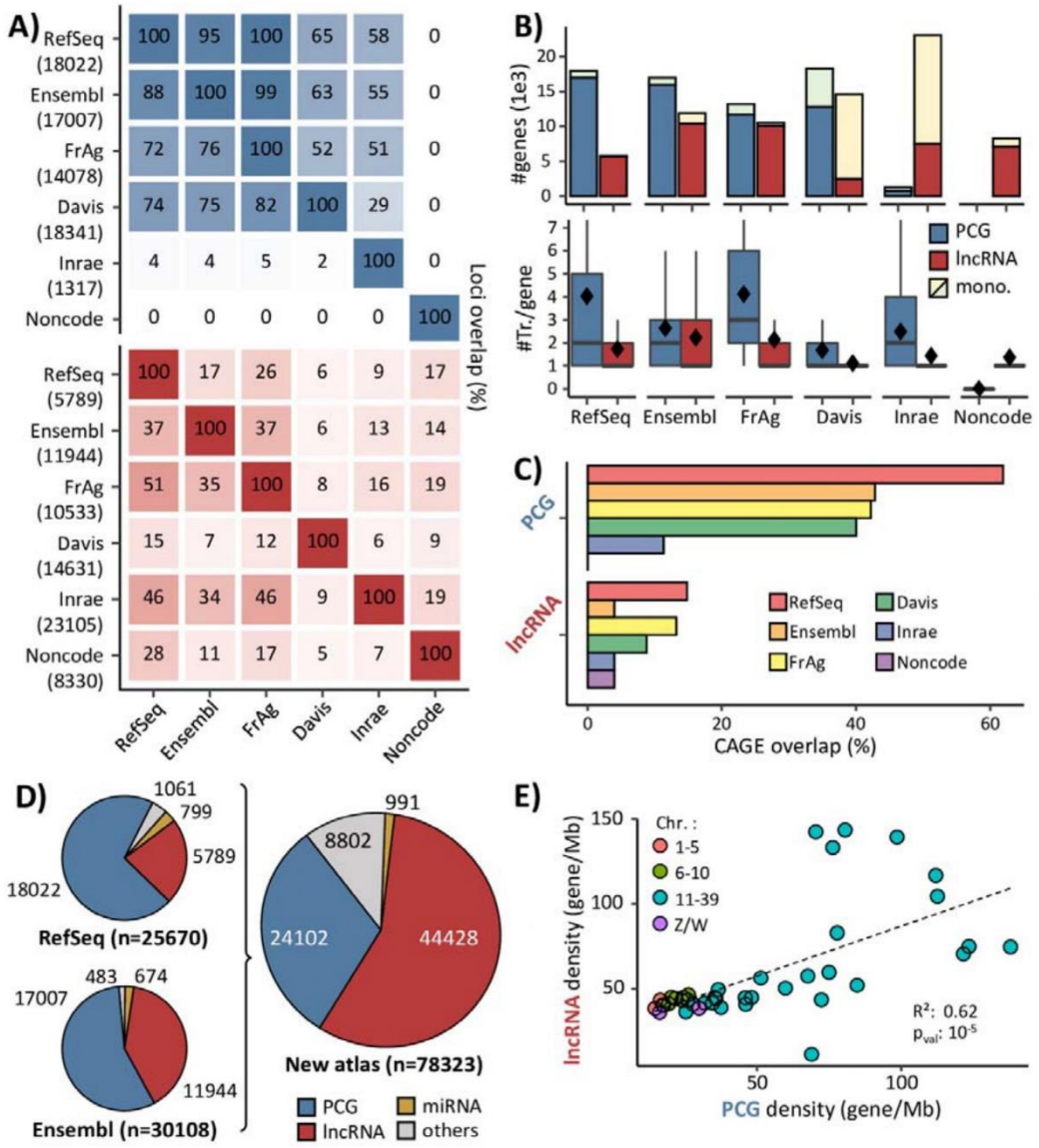


Figure 1. Characteristics of the gene-enriched annotation and its component sources

(a) % of overlapping PCGs (blue) and lncRNAs (red) having at least 1 bp in common for exons on the same strand, between the databases. % in upper triangle refer to x-axis. The number of loci per database is indicated in line. (b) Number of PCGs and lncRNAs and number of transcripts per gene by databases. Diamonds indicate the average value. mono.: monoexonic. (c) % of PCG and lncRNA TSSs overlapping a CAGE peak within +/- 30bp. (d) Proportion of gene biotypes in the "RefSeq" and "Ensembl" reference databases and in the enriched genome annotation. (e) Correlation between lncRNA density and PCG density across the chicken macro-, meso-, micro- and sexual chromosomes.

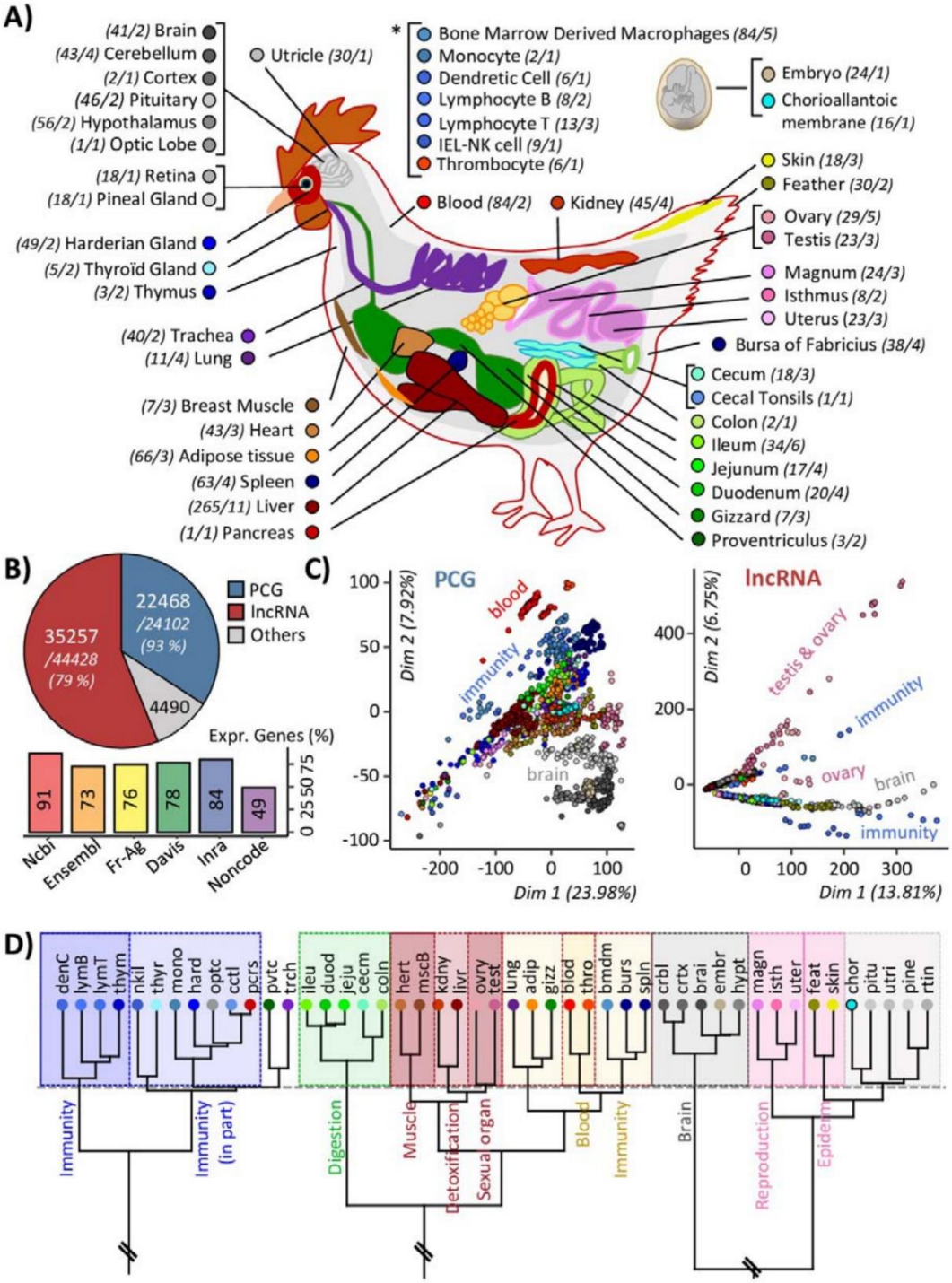


Figure 2. Gene expression across 47 chicken tissues.

(a) Illustration of the 47 tissues used for gene expression. Numbers in parentheses correspond to the number of samples and the number of constitutive datasets. Corresponding colours are

indicated in the adjacent circles. Full tissue names are available in Sup. Table. 11 (b) Top: Numbers of PCGs (blue) and lncRNAs (red) considered as expressed applying a normalized expression threshold of 0.1 TPM and TMM. Bottom: % of expressed genes according to the constitutive sources of the enriched annotation. (c) Principal component analysis based on the gene expression of expressed PCGs (left) and lncRNAs (right). (d) Hierarchical clustering of the expressed genes for the 47 tissues and performed using "1-Pearson correlation" distance and "ward" aggregation criteria.

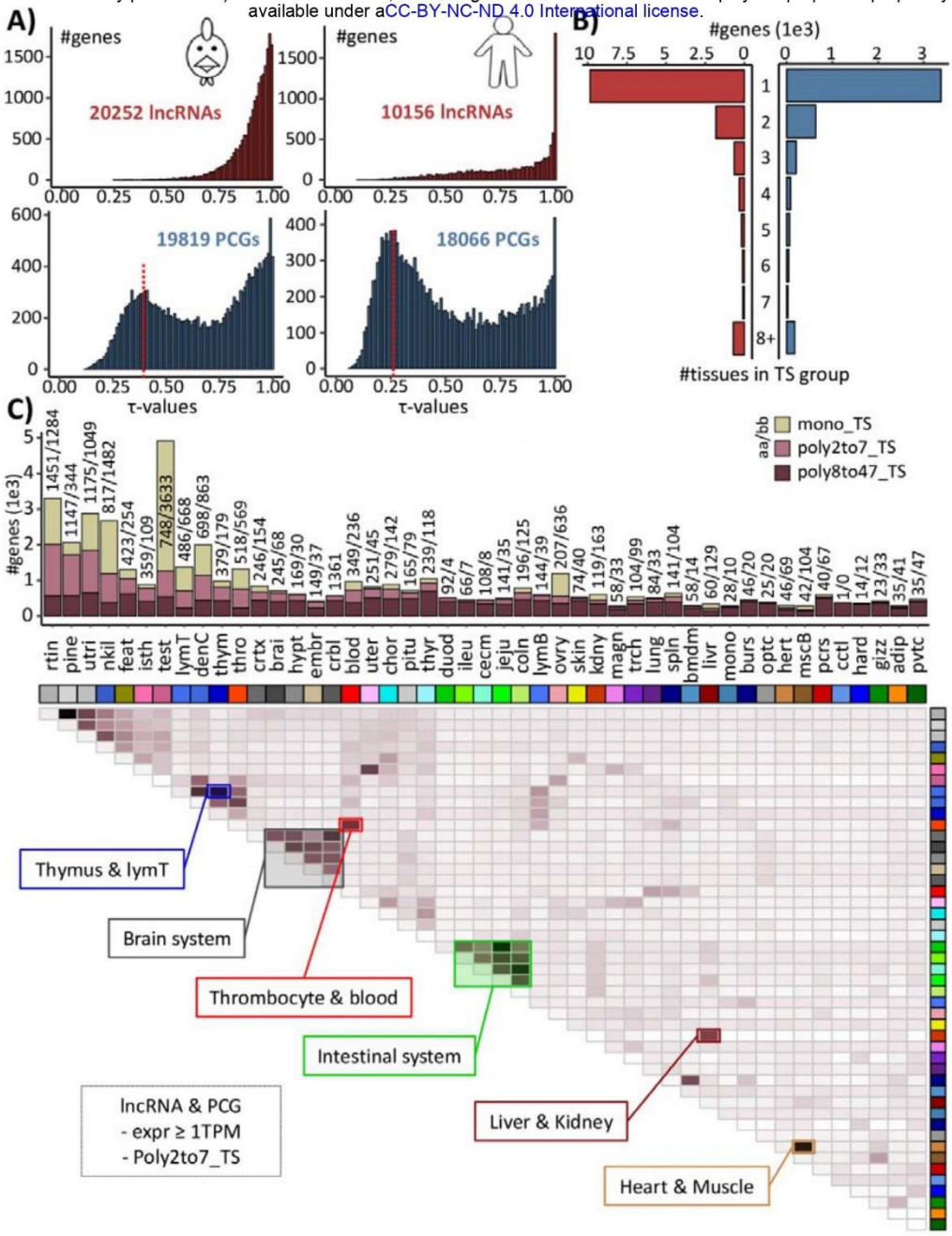


Figure 3. Tissue specificity across 47 chicken tissues.

(a) Distribution of τ values for lncRNAs (red) and PCGs (blue) with an expression ≥ 1 TPM for chicken (left) and human (right). The red dotted line indicates the first local maximum associated to ubiquitous genes. (b) Distribution of lncRNAs (red) and PCGs (blue) with an expression ≥ 1 TPM according to the number of tissues for which the gene is considered as tissue-specific. (c) Number of mono_TS (light brown), poly2to7_TS (pink-brown) and poly8to47_TS (dark brown), tissue-specific (TS) genes per tissue (top) and clustered heatmap based on pairwise association (bottom). Full tissue names are available in Sup. Table. 11.

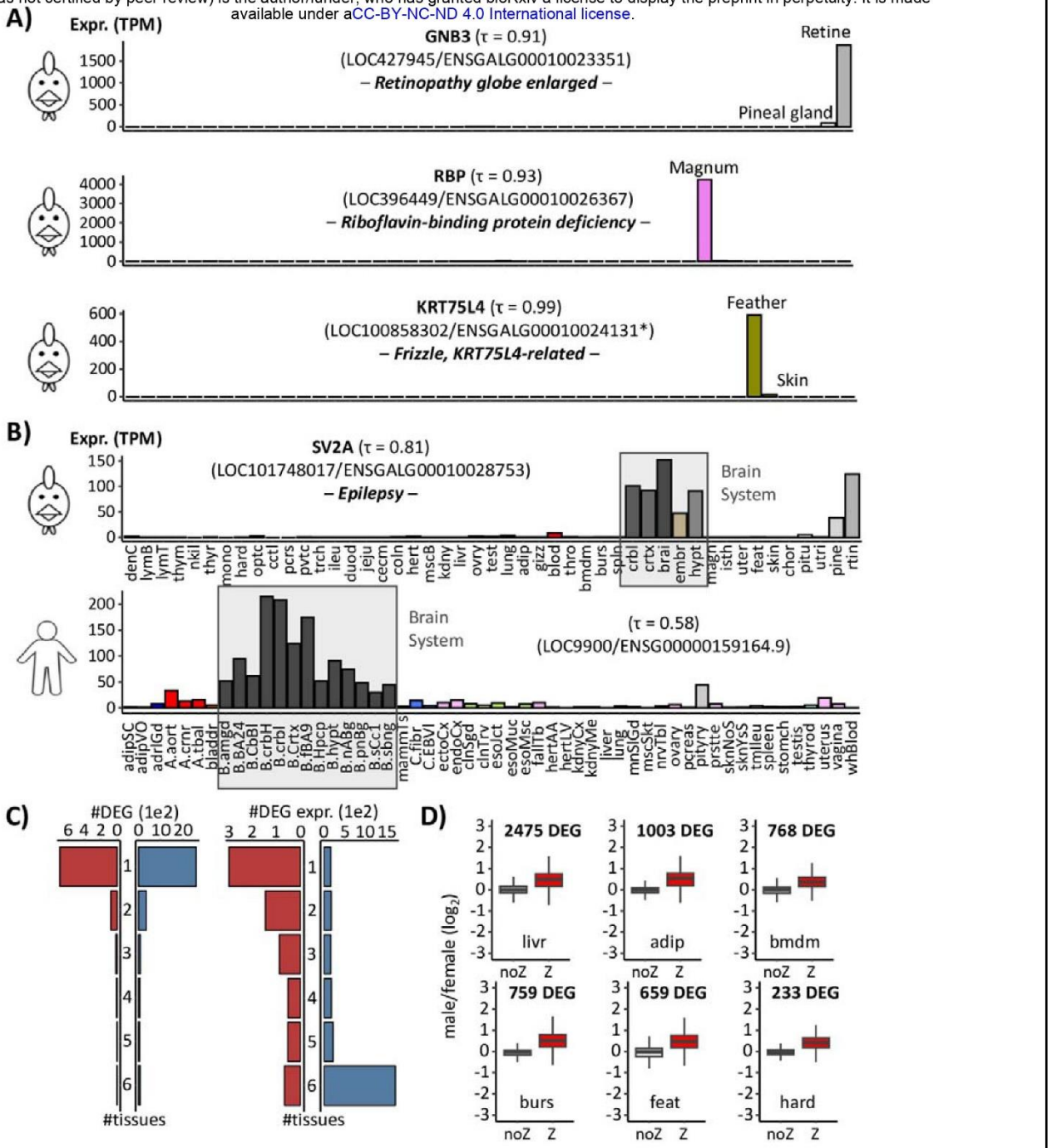
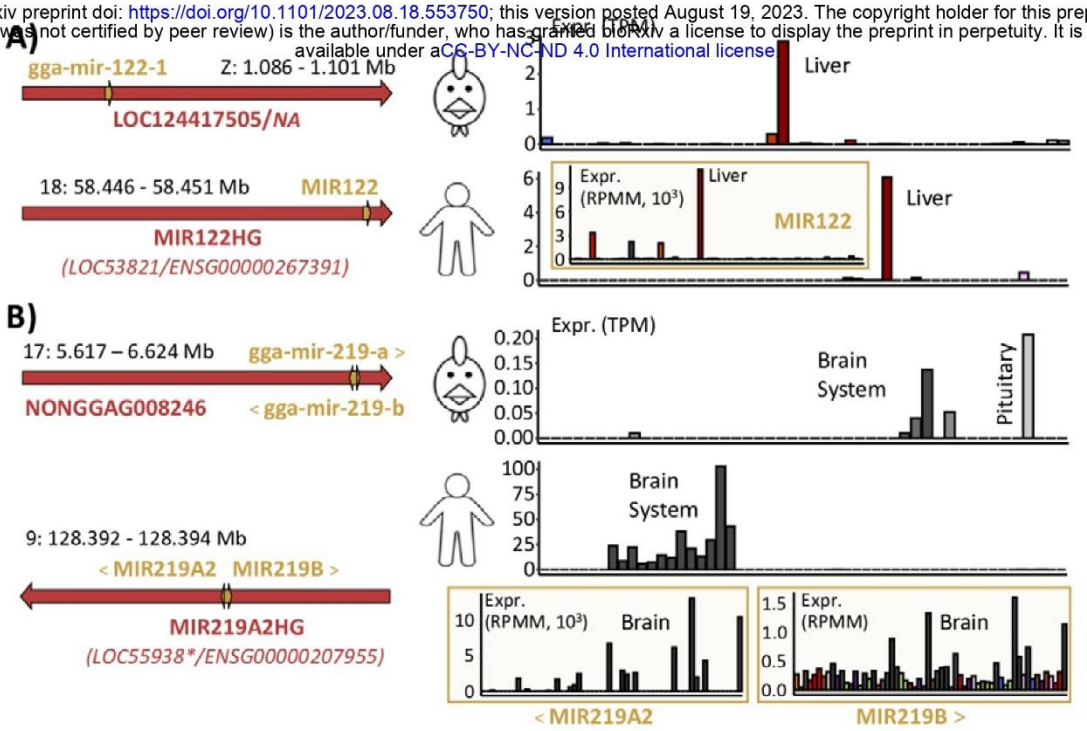


Figure 4. Illustrative cases of gene expression interest for functional analyses.

(a) Expression profiles in TPM of 3 tissue-specific genes associated with a Mendelian trait in chicken: GNB3 retina-specific (top), RBP magnum-specific (middle), and KRT75L4 (bottom) feather-specific (bottom right). Both "RefSeq" and "Ensembl" gene identifiers are provided. (*) indicates that the gene identifier equivalence is not provided by BioMart but was found by overlap between the two reference genome annotations. (b) Expression profile of SV2A in

TPM in chicken (top) and human (bottom). Full tissue names for chicken are available in Sup. Table. 11. The 53 human GTEx tissues are ordered, abbreviated and coloured as indicated in the Sup. Table 12. (c) Left: Number of differentially expressed genes (DEG) shared between the 6 tissues. Right: Number of genes identified as DEG in at least one tissue and considered as expressed across the 6 tissues. (d) log2(Fold Change) of differentially expressed genes (DEGs) between sexes for 6 tissues and excluding the "Z" chromosome.



C)

lncRNA:PCG pairs	Intergenic				Genic	
	Convergent	*Divergent*	*SS. up*	*SS. down*	*Antisense*	*Sense*
Total pairs: 33907	**2728**	**4072**	**3382**	**4561**	**12537**	**6627**
≤ 5kb	1280 (47%)	2263 (56%)	1557 (46%)	1983 (43%)		
> 5kb	1448 (53%)	1809 (44%)	1825 (54%)	2578 (57%)		
median distance (bp)	*5755*	*3572*	*6098*	*6649*		
Co-expressed pairs: 2331 (6.9%)	**46 (1.7%)**	**317 (7.8%)**	**178 (5.2%)**	**577 (12.6%)**	**439 (3.5%)**	**774 (11.7%)**
≤ 5kb	23 (1.8%)	263 (11.6%)	106 (6.8%)	390 (19.7%)		
> 5kb	23 (1.6%)	54 (3.0%)	72 (3.9%)	187 (7.3%)		
median |ρ|	*0.11*	*0.18*	*0.14*	*0.22*	*0.13*	*0.19*

PCG:PCG pairs	Intergenic				Genic	
	Convergent	*Divergent*	*SS. Up*	*SS. Down*	*Antisense*	*Sense*
Total pairs: 20656	**2525**	**3471**	**3067**	**3348**	**5246**	**2999**
≤ 5kb	1718 (68%)	2249 (65%)	1951 (64%)	2249 (67%)		
> 5kb	807 (32%)	1016 (35%)	1116 (36%)	1099 (33%)		
median distance (bp)	*3130*	*2134*	*3566*	*3185*		
Co-expressed pairs: 3375 (16.3%)	**257 (10.2%)**	**834 (24.0%)**	**622 (20.3%)**	**732 (21.9%)**	**450 (8.6%)**	**480 (16.0%)**
≤ 5kb	181 (10.5%)	663 (29.5%)	425 (21.8%)	526 (23.4%)		
> 5kb	76 (9.5%)	171 (16.8%)	197 (17.8%)	206 (18.7%)		
median |ρ|	*0.31*	*0.4*	*0.38*	*0.38*	*0.19*	*0.27*

Figure 5. Genomic configuration and co-expression using the extended annotation.

(a-b) Conservation of the genomic configuration (left) and expression profile in TPM (right), between the 47 chicken tissues (top) and the 53 human GTEx tissues (bottom). Mir expression is shown in the yellow rectangle. (a) MIR122HG gene, host of mir122 identified in human, has an equivalent locus in the chicken reference databases but is unnamed. (b) MIR219A2HG gene, host of mir219a2 and mir219b identified in human, has an unnamed equivalent locus in

the extended chicken annotation but not in the reference databases. (*) indicates the old gene identifier for the human "RefSeq" database which is no longer used, the gene model being removed. (c) Classification of lncRNAs (top) and PCGs (bottom) according to their closest PCG and co-expression. SS. up: Same strand up, SS. down: same strand down.

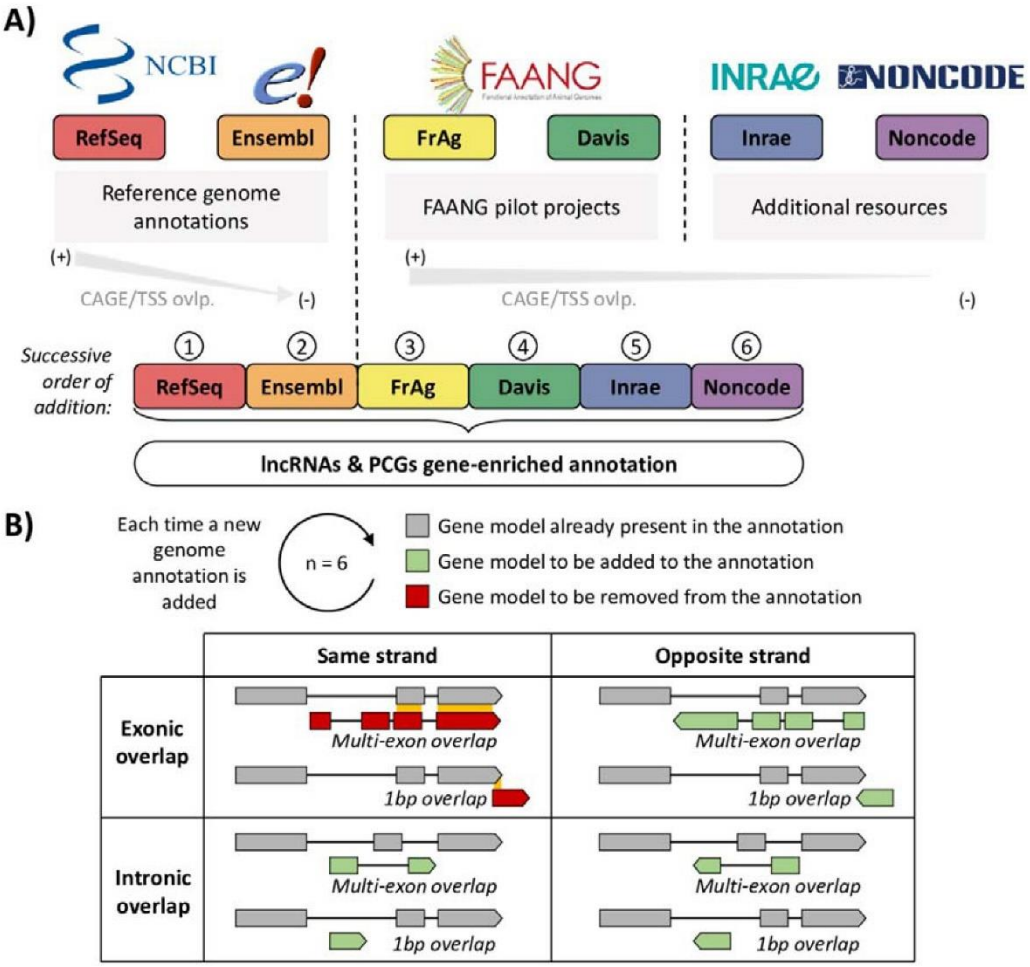
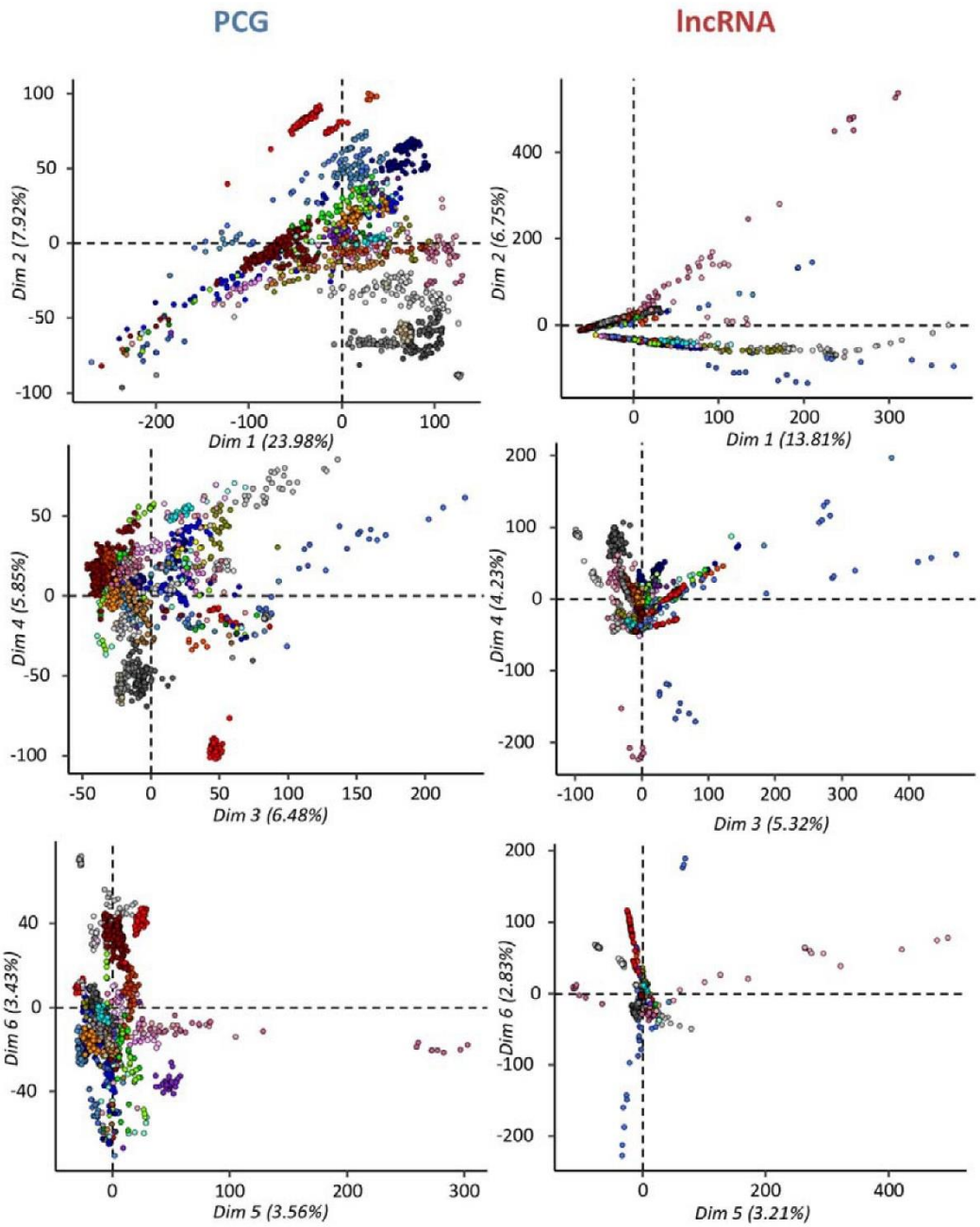


Figure 6. Gene-enriched annotation construction.

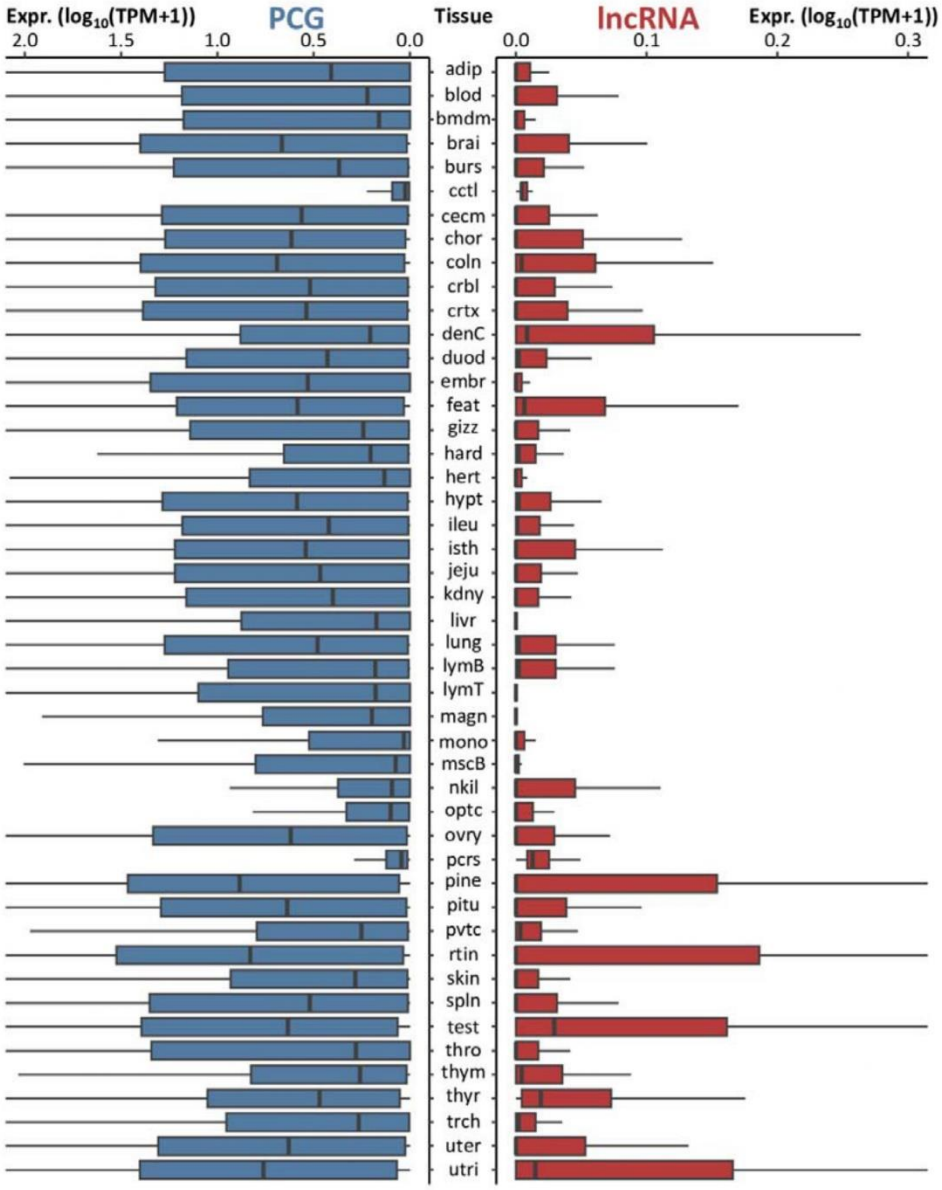
(a) Origin and order of the successive addition of the 6 genome annotations used to build the gene-enriched annotation. TSS: Transcription Start Site of the transcript models, ovlp.: overlap. (b) Aggregation rules applied each time a new genome annotation is added with respect to the pre-existing gene models.

SUPPLEMENTARY FIGURES



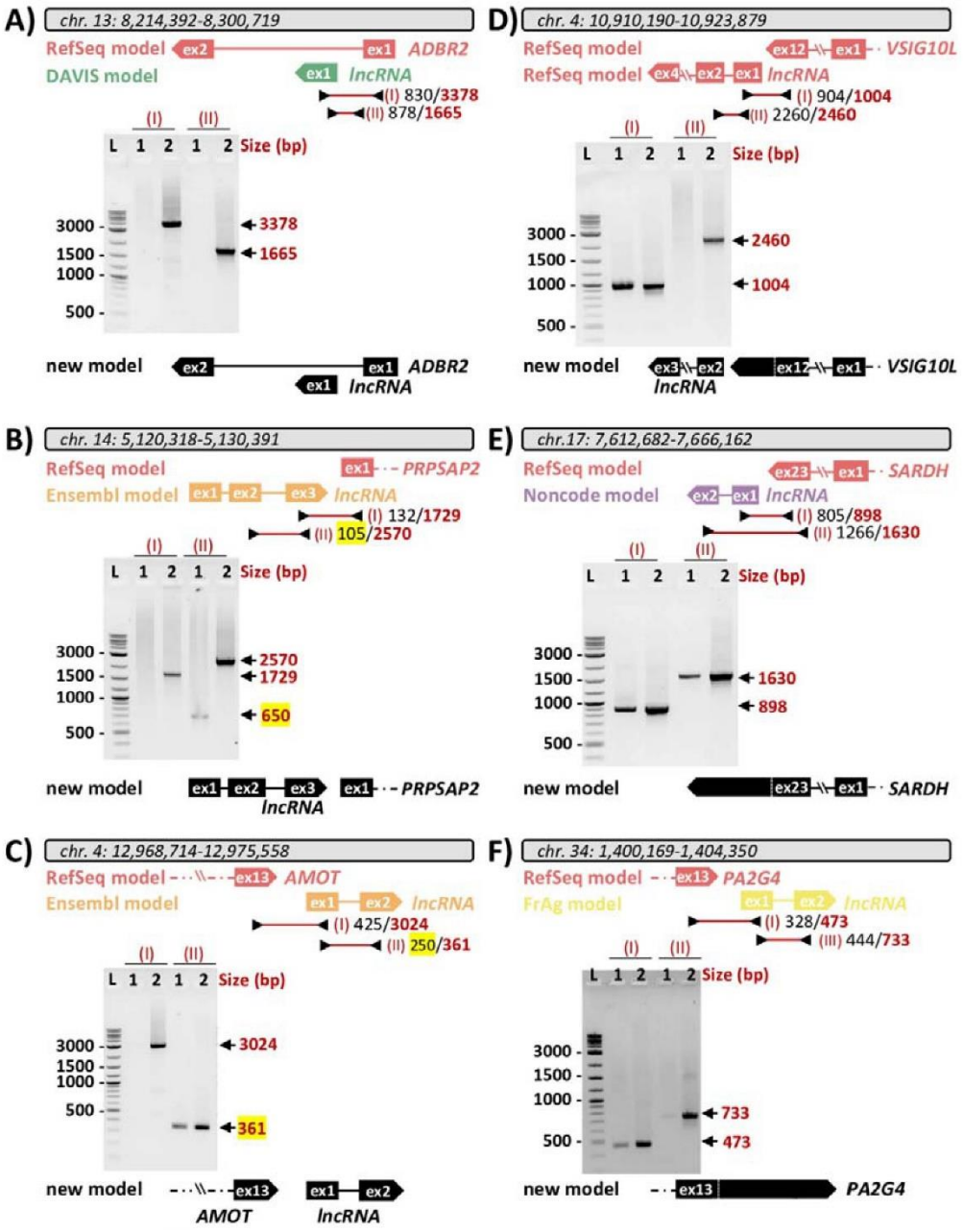
Sup. Figure 1. Principal component analysis based on gene expression of expressed PCGs and lncRNAs.

The factorial plans for axes 1:2, 3:4 and 5:6 are provided. Colours and associated tissues are available in Sup. Table. 11.



Sup. Figure 2. Distribution of PCG (blue) and lncRNA gene expression in $\log_{10}$(TPM+1) in chicken for the 47 tissues.

Full tissue names for chicken are available in Sup. Table. 11.



Sup. Figure 3. Reliability of six lncRNAs in same-strand configuration of a PCG tested by PCR.

Left: lncRNAs considered as independent loci from the (a) DAVISGALG000044072/ADBR2, (b) ENSGALG00010022678/PRPSAP2, and (c) ENSGALG00010016012/AMOT lncRNA:PCG pairs. Right: lncRNAs considered as extension of the PCG from the (d) LOC121113202/VSIG10L, (e)

NONGGAG001811/SARDH, and (d) FRAGALG000000006896/PA2G4 lncRNA:PCG pairs. The upper part of each panel represents the relative position of the constituent genes of the lncRNA:PCG pair as identified on the enriched atlas. The lower panel shows the constituent genes of the lncRNA:PCG pair based on the PCR results. The letters/numbers above each gel correspond to: L: ladder; 1: PCR using cDNA; 2: PCR using genomic DNA (gDNA). The roman numerals refer to the PCR primer pair used which are indicated in the upper part with the predicted size for cDNA and gDNA. Arrows next to the band indicate the observed size of the amplified fragment in relation to what was predicted.

SUPPLEMENTARY TABLES

Sup. Table 1. Gene annotation with genomic and functional information/features for gene models of the enriched-atlas including the orthology, the expression, the tissue-specificity, the classification of gene models with the closest PCG or lncRNA, GO terms but also identifiers equivalence between the two reference genome annotations "RefSeq" and "Ensembl". Also available at www.fragencode.org/lnchickenatlas.html with the corresponding genome annotation (*.gtf*).

Sup. Table 2. Characteristics of the gene models included in each genome annotation used to build the enriched-annotation. (a) Size and number of genes, transcripts, exons and their associated proportions for lncRNAs, PCGs and all gene models. (b) Number of lncRNAs and PCGs supported by one ("1tr") or more ("Xtr") transcripts and with one ("1ex") or more ("Xex") exons. Transcripts classified as multi-exonic but with only one exon longer than 50bp are considered as "False Multi- exonic" ("FM"). (c) Number and types of biotypes indicated in each database.

Sup. Table 3. Number of genes and their associated biotypes successively added per database used to build the enriched-annotation.

Sup. Table 4. Project accession numbers and number of samples used to quantify the gene expression across the 47 tissues composing the atlas.

Sup. Table 5. Number of expressed and tissue-specific PCGs and lncRNAs across the 47 tissues for an expression threshold of 0.1 and 1 TPM. mono_TS: genes specific to a single tissue, poly2to7_TS and poly8to47_TS: genes specific to a group of n tissues with $n \leq 7$ and $n > 7$ respectively. Full tissue names for chicken are available in Sup. Table. 11.

Sup. Table 6. Genes related to a known Mendelian trait or disorder ("Phene") obtained from the OMIA resource. The hypothetical tissue in which the causative gene/variant is likely to have an effect is indicated in the "ExpectedTissue" column. For each gene, its name ("GeneName"), its genes identifier in "RefSeq" ("GeneId") and in "Ensembl" both by BioMart ("GeneId_BiomartEnsEq") and by overlap ("GeneId_OvlpEnsEq") are provided according to the GRCg7b assembly.

Sup. Table 7. List of differentially expressed genes (DEGs) between sexes (male/female) detected in the liver (livr), adipose tissue (adip), bone marrow-derived macrophages (bmdm), bursa of Fabricius (burs), feather (feat), and the Harderian gland (hard). For "bmdm", the analysis was conducted on two independent projects and the union of DEGs was used. For each gene, its name ("GeneName"), its genes identifier in "RefSeq" ("GeneId") and in "Ensembl" both by BioMart ("GeneId_BiomartEnsEq") and by overlap ("GeneId_OvlpEnsEq") are provided according to the GRCg7b assembly.

Sup. Table 8. Equivalence table of the gene identifiers from our previous annotation in galgal5 and GRCg6a to the one in GRCg7b. Two types of list are provided: *i)* an equivalence gene by gene with the coordinates in both assembly; *ii)* an equivalence only with gene identifiers collapsed considering GRCg7b as the reference.

Sup. Table 9. Numbers of lncRNAs and PCGs according to their configuration with their closest PCG and their genome annotation origin.

Sup. Table 10. Priorization of the gene biotypes applied when gathering the different genome annotations.

Sup. Table 11. Names, abbreviations and colours of the 47 chicken tissues.

Sup. Table 12. Names, abbreviations and colours of the 53 human GTEx tissues.

Sup. Table 13. Primers sequences and corresponding annealing temperature used for PCR analysis of lncRNA:PCG pairs in same strand configuration.

POSTER

A lncRNA GENE-ENRICHED ATLAS FOR THE GRCg7b CHICKEN GENOME AND ITS FUNCTIONAL ANNOTATION ACROSS 47 TISSUES

F. Degalez[1] & S. Lagarrigue[1]

fabien.degalez@inrae.fr

sandrine.lagarrigue@institut-agro.fr

[1]PEGASE, INRAE, Institut Agro, 35590 Saint Gilles, France

CONTEXT

While protein-coding genes (PCG) are relatively well known in livestock genomes, long non-coding RNA (lncRNA) gene loci, known to be regulatory genes, are more poorly described and differ greatly between reference genome annotations. Even if new expression data contribute to improve their identification, their low expression and high context-specificity remain a challenge.

For the chicken, in 2022, the new GRCg7b chicken genome assembly with its associated genome annotations have been released.

OBJECTIVES

- Provide an annotation of the chicken genome according to the GRCg7b assembly.
- Integrate the two reference annotations including "NCBI-RefSeq" (Ncbi) and "EMBL-EBI Ensembl/GENCODE" (Ensembl).
- Ease the joint use of gene models from Ncbi & Ensembl and the switch between the galgal5, GRCg6a and GRCg7b chicken assemblies.
- Increase the number of lncRNA identified using additional resources from multi-tissue projects or specialist databases.
- Provide a genomic and functional annotation for the community working on gene expression (lncRNAs and/or PCGs) to elucidate, for example, the molecular mechanisms linking non-coding variants and relevant phenotypes.

MATERIALS & METHODS

Fig. 1: Origin and aggregation rules used to build the gene-enriched annotation

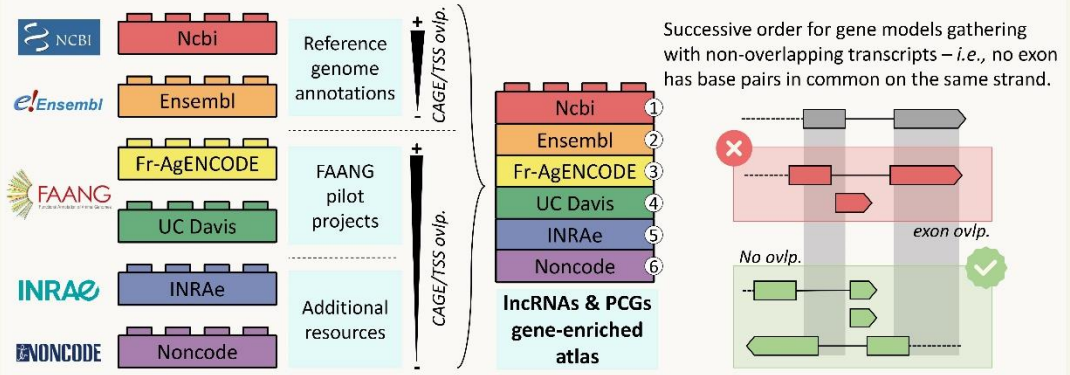
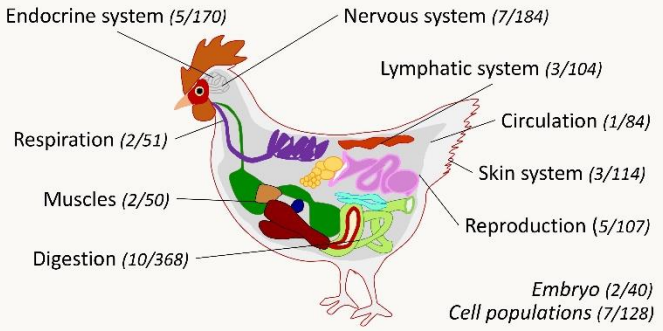


Fig. 2: A diversity of 47 tissues used for gene expression



(47 tissues / 1400 samples) from 36 datasets

RESULTS

Fig. 3: Number of genes in the enriched atlas

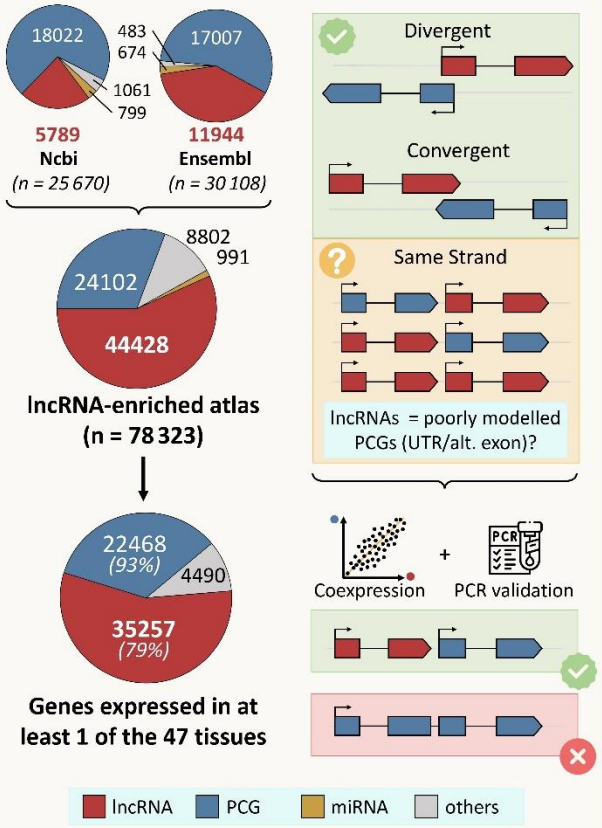
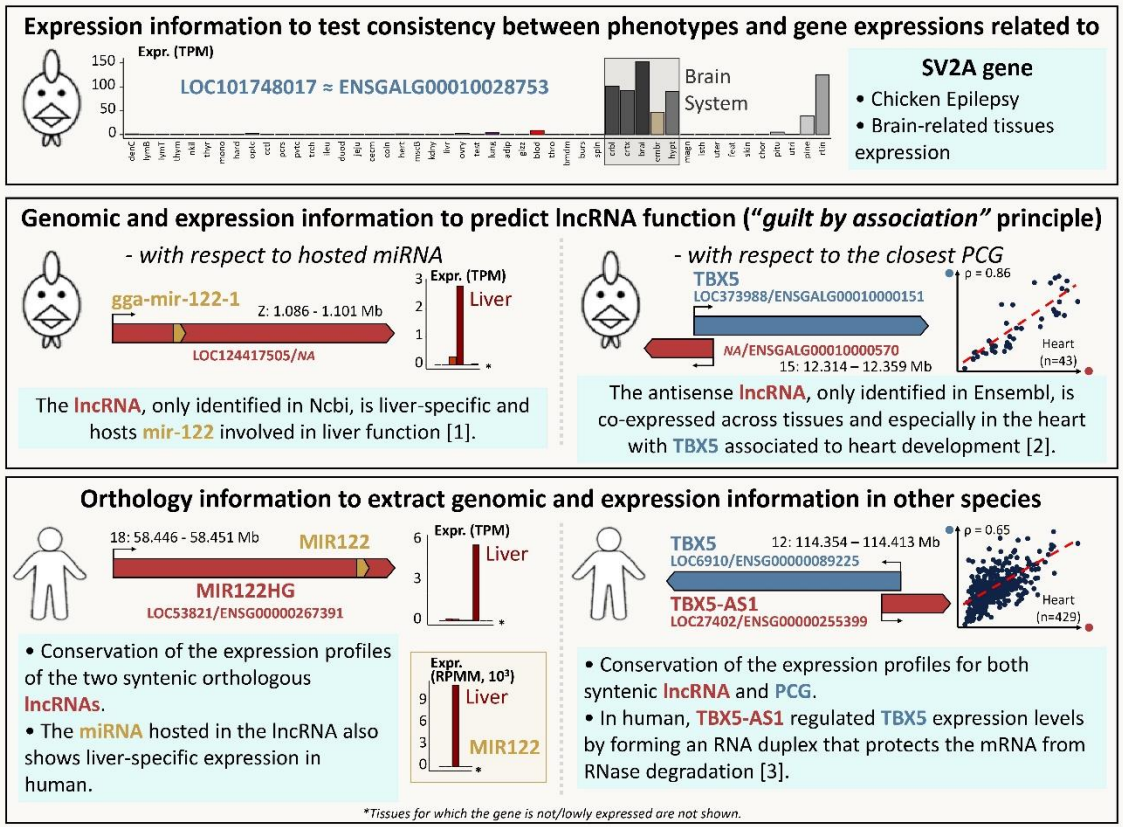
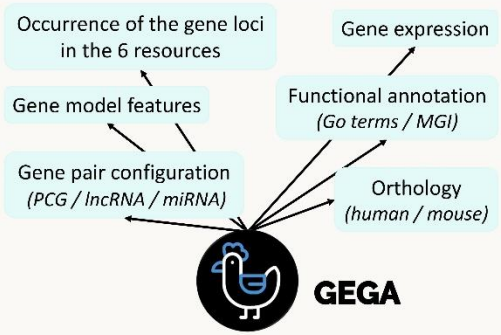


Fig. 4: Functional gene annotation, a resource for a variety of analyses





WEBSITE

To help in performing studies such as QTL, eQTL or differentially expressed genes analyses, all these data are easily accessible via a user-friendly online tool called **GEGA (Gallus Enriched Gene Annotation)** enabling efficient data analysis thanks to a multitude of modules and an ergonomic design. All results and a multitude of graphical representation can be simply produced and downloaded, ranging from basic gene information to detailed visualization of expression and co-expression.

gega.sigenae.org

Acknowledgments: This project is funded by the European Union's Horizon 2020 research and innovation program under grant agreement N°101000236 and by ANR CE20 under 'EFFICACE' program.

References: [1] Bandiera S, et al. miR-122--a key factor and therapeutic target in liver disease. J Hepatol. 2015;62:448–57; [2] Steimle JD, Moskowitz IP. TBX5: A Key Regulator of Heart Development. Current Topics in Developmental Biology. Academic Press; 2017. p. 195–221. [3] Ma J, et al. Hypermethylation-mediated down-regulation of lncRNA TBX5-AS1:2 in Tetralogy of Fallot inhibits cell proliferation by reducing TBX5 expression. Journal of Cellular and Molecular Medicine. 2020;24:6472–84.

1.2. GEGA : un outil en ligne facilitant l'exploration des annotations générées *(Résumé d'article)*

1.2.1. Contexte et objectifs

Ces dernières années, des progrès significatifs ont été réalisés dans l'annotation du génome de la poule, notamment grâce à l'émergence des technologies de séquençage haut débit et aux efforts continus de projets de recherche collaboratifs tels que FAANG [196], NONCODE [112] ou le projet ChickenGTEx [405] (partie du projet FarmGTEx [457]). Cependant, pour assurer l'annotation la plus complète possible, un levier est de croiser les annotations de sources multiples en utilisant une variété de données et de *pipelines* d'analyse, et plus particulièrement celles de NCBI-RefSeq et EMBL-EBI Ensembl/GENCODE qui sont largement reconnues comme des ressources de références fiables [103, 124, 125]. Comme précédemment démontré pour le dernier assemblage GRCg7b du génome de la poule (voir Introduction §1.2 ; Résultats §1.1), les modèles de gènes entre ces deux annotations de référence sont assez similaires en termes de PCG avec environ 90 % des loci partagés, même si les modèles de transcrits sous-jacents restent différents. Concernant les lncRNA, gènes régulateurs de l'expression et contribuant à une variété de processus biologiques, ils sont plus difficiles à identifier en raison de leurs caractéristiques spécifiques aux conditions (tissulaires et temporelles). Ces modèles de gène ne partagent alors qu'approximativement 25 % de leurs loci entre les annotations RefSeq et Ensembl [125]. Cependant, même si certaines initiatives, comme le projet MANE pour l'humain [429], visent à déterminer des modèles de gènes communs entre ces annotations, à ce jour et à notre connaissance, une annotation résultant de l'union de ces bases de données de référence n'est actuellement pas accessible de manière systématique et facile. Cependant, la disponibilité d'une annotation du génome aussi complète que possible apparait essentielle pour la communauté scientifique et notamment chez la poule, où les gènes et en particulier les lncRNA ne sont encore que partiellement connus. Dans ce contexte, nous avons récemment publié un atlas enrichi en gène et basé sur le dernier assemblage GRCg7b de la poule. Cet atlas intègre les annotations de référence RefSeq et Ensembl ainsi que des ressources supplémentaires (voir Résultats §1.1) [125]. Cette annotation se compose au final de 78 323 modèles de gènes, dont 24 102 PCG et 44 428 lncRNA, augmentant ainsi considérablement le nombre de gènes fournis par chaque ressource séparément. D'autre part, cet atlas est accompagné d'informations complémentaires

permettant d'approfondir les aspects fonctionnels, notamment via les profils d'expression et de co-expression au travers d'une collection de 47 tissus provenant de 36 ensembles de données et 1400 échantillons. Ces données permettent, par exemple, de s'intéresser à la tissu-spécificité ou encore aux variations d'expression en fonction du sexe ou de l'âge. En parallèle, des données en lien avec l'orthologie avec l'humain et la souris ou des classifications positionnelles entre les gènes sont également disponibles (voir *Graphical Abstract*). Pour faciliter l'accès et l'exploration de toutes ces informations, nous avons développé un outil facilement accessible en ligne appelé GEGA (*Gallus Enriched Gene annotation*). GEGA s'adresse à toutes les communautés, qu'il s'agisse de celles intéressées par des régions génomiques associées à des phénotypes d'intérêt (QTL) ou de celles travaillant directement sur l'expression des gènes et intéressées par un gène spécifique ou une liste de gènes. Les résultats peuvent notamment être téléchargés facilement, que ce soit via des tables choisies par l'utilisateur ou via la multitude de représentations graphiques disponible.

1.2.2. Résultats

GEGA est un outil en ligne (https://gega.sigenae.org/) permettant l'exploration d'un atlas de 78 323 modèles de gènes comprenant à la fois des annotations génomiques et fonctionnelles selon le dernier assemblage du génome GRCg7b de la poule. L'outil se divise en deux éléments interconnectés (voir Figure 1) :

- Une table interactive pour visualiser les différentes caractéristiques des gènes.
- Une interface graphique permettant d'explorer à la fois les expressions et co-expression pour les modèles de gènes en fonction de différentes conditions (inter-tissus / intra-tissus ; sexe / âge), mais aussi les informations sur les modèles de gènes précédemment sélectionnés dans le tableau interactif.

Concernant la table interactive, deux éléments permettent de naviguer au travers des différentes données, le sélecteur de vue ainsi que le module de filtration.

Avec le sélecteur de vue (Figure 1 – Panel A), l'utilisateur a la possibilité soit i) de sélectionner les colonnes qu'il désire afficher (via « Customize »), soit ii) d'utiliser les vues prédéfinies et qui suivent les principales catégories d'informations disponibles dans GEGA. Dans ce deuxième cas, les trois premières colonnes, qui sont toujours identiques, comprennent l'identifiant du gène et les noms courts et longs. Les vues préconfigurées sont les suivantes :

- « *Default* » fournissant l'origine du modèle génique, les informations positionnelles et le biotype ;
- « All *Naming* » affichant les identifiants disponibles pour un modèle de gène, en particulier Ensembl et RefSeq, et également selon les assemblages précédents (galgal5 et GRCg6a) ,
- « *All Functional* », énumérant les termes GO (*Gene Ontology*) et les phénotypes connus associées pour chaque gène et pour leurs orthologues ;
- « *Expression & Tissue-specificity* », indiquant les valeurs d'expression (en TPM) dans chacun des 47 tissus, le premier et le second tissus les plus exprimés ainsi que des informations sur la spécificité tissulaire ;
- « *Orthology* », soulignant les relations d'orthologie et les équivalents géniques avec l'humain et la souris (à ce jour, uniquement pour les PCG) ;
- « *Gene Structure* », détaillant la structure du gène en termes de transcrit, d'exons et d'introns ;
- « *FEELnc* » indiquant le modèle de gène le plus proche (PCG/lncRNA/miRNA) et la configuration au niveau du gène et du transcrit, ainsi que les données de coexpression au niveau du gène ;
- « *Repetability* », fournissant des informations sur la reproductibilité de la représentation des loci dans l'ensemble des six ressources utilisées.

Avec le module de filtration (Figure 1 – Panel B1), l'utilisateur peut appliquer des filtres sur l'ensemble des colonnes afin d'extraire facilement un ensemble spécifique de gènes. À l'aide de "fonctions personnalisées" prédéfinies, il est possible de rechercher dans plusieurs colonnes en un seul filtre. À ce jour, les fonctions "*Function all naming search*" et "*Function all functionnal annot. search*" permettent de filtrer respectivement sur toutes les colonnes relatives à la dénomination (noms et identifiants de gènes provenant à la fois de Refseq et d'Ensembl) ou aux termes GO. Pour un filtre donné, le délimiteur virgule (",") peut être utilisé pour indiquer plusieurs possibilités (opérateur *OR*), par exemple, l'application du filtre « *gnSimpleBiotype - contains - lnc,pcg* » entraînera une sélection à la fois des lncRNA et des PCG. De même, « *Function all naming search - contains - LOC417220, ENSGALG00010025549, ACACA* » aboutira à une sélection des trois gènes mentionnés, même si le type de nomination utilisé varie. Notons que chaque ligne correspond à un critère et qu'ainsi, un ensemble de critères est considéré comme une intersection de conditions (opérateur *AND*). Notons que même si l'expression de chaque tissu n'est pas disponible individuellement dans le tableau interactif, il est possible de filtrer l'expression du gène du tissu à l'aide du filtre « *[tissueName]_expr* ».

Afin de répondre à des usages spécifiques (voir les « Focus », juste après), trois modules ont été conçus pour générer facilement les filtres nécessaires à l'utilisation prévue : *i)* « *Gene list* » pour saisir rapidement une liste de gènes ; *ii) « Define a region »* qui crée une région autour d'un gène ou d'une position et qui peut être utilisée, par exemple, pour l'analyse des QTL ; *iii) « Expression threshold »* permettant d'appliquer facilement une combinaison de filtres basés sur des critères d'expression.
Finalement, les résultats d'annotation ou d'expression peuvent être exportés sous la forme de deux fichiers (*.tsv ou .xlsx)* et les gènes d'intérêt peuvent être facilement sélectionnés via une copie dans le presse-papier des positions ou des identifiants.

Concernant l'interface graphique, il est tout d'abord possible de générer des *pie charts*, des *boxplots* et des *scatter plot* pour les variables catégorielles et numériques et pour les gènes précédemment sélectionnés via la table interactive. Dans un second temps et en raison de la variabilité des données, les profils d'expression peuvent être observés à différents niveaux. Alors que les profils d'expression peuvent être analysés en comparant les 47 tissus (*inter-tissue ; Figure 1 – Panel H*), la variation entre les individus et les projets peut également être observée au sein d'un tissu spécifique (*intra-tissu ; Figure 1 – Panel G*). De même, après avoir sélectionné un tissu d'intérêt, un projet peut être sélectionné pour visualiser l'expression en fonction de critères tels que le sexe ou l'âge *Figure 1 – Panel I*. Dans chacun des cas, les *boxplot* et les diagrammes de coexpression (*scatter plot*) sont disponibles et fonctionnent de manière similaire. Pour les *boxplot*, un gène ou une liste de gènes peut être fourni. Différentes options d'affichage sont disponibles, incluant l'unification des axes d'expression (0 à la valeur maximale), l'aperçu des points individuels, ou alors la classification des tissus par abréviation ou en fonction de la proximité d'expression obtenue par classification (voir Résultats §1.1) [125]. Pour les diagrammes de coexpression, un identifiant de gène de référence est d'abord fourni, suivi d'un ou de plusieurs gènes cibles. Les différents graphiques peuvent ensuite être ordonnés par corrélation ou filtrés par un seuil de corrélation. La co-expression pouvant suivre différentes lois, des tracés linéaires ou logarithmiques sont disponibles.
Pour examiner l'expression *inter-tissu*, il est possible de générer soit un *boxplot*, soit un *barplot* de l'expression pour les 47 tissus. Dans notre cas, ces deux visualisations correspondent en réalité à des analyses différentes. Les *boxplot* considèrent tous les échantillons de tous les

projets ensemble, en traçant toutes les données disponibles. Les *barplot* calculent d'abord l'expression médiane pour chaque projet, puis la médiane de ces médianes, expliquant le manque de résolution au niveau de l'échantillon et la variabilité des médianes entre les deux visualisations.

Spécifiquement aux analyses avec des critères tels que le sexe ou l'âge, il convient d'abord de sélectionner un projet d'intérêt. Cela permet de garantir des comparaisons fiables et cohérentes entre les sous-groupes en fonction des facteurs étudiés. Les *boxplots* peuvent être générés pour un tissu spécifique ou jusqu'à quatre tissus. Finalement, toute figure générée peut être téléchargée au format PNG, JPEG, SVG ou PDF via le menu dédié en haut à droite du tracé. Les données utilisées pour réaliser les graphes peuvent être téléchargées sous forme de fichiers *.tsv ou .xlsx* pour permettre à l'utilisateur de régénérer les figures selon ses propres souhaits.

Afin de montrer les possibilités d'usage de GEGA, trois exemples d'utilisations courantes sont présentés.

- **Focus sur un gène spécifique** : (*e.g.*, TBX5 ; Figure 2) : TBX5, un facteur de transcription [458], joue un rôle essentiel dans la morphogenèse cardiaque des vertébrés. La surexpression de TBX5 dans les cœurs d'embryons de poussins a montré que TBX5 inhibe la croissance du myocarde et participe ainsi à la modulation de la croissance et du développement cardiaques chez les vertébrés [459]. En utilisant GEGA, ce gène peut ensuite être exploré plus en profondeur, selon cette démarche :

1) Analyse de l'expression au travers des 47 tissus – TBX5 est expression-spécifique du cœur ;
2) Analyse des termes fonctionnels – termes associés au cœur ;
3) Observation de la configuration génomique avec les gènes voisins – Présence d'un lncRNA en configuration antisens ;
4) Recherche d'orthologie avec l'humain et la souris – TBX5 a un PCG orthologue dans les deux espèces ce qui a servi pour la vérification manuelle de la conservation du lncRNA antisens dans les deux espèces.
5) Analyses de l'expression du lncRNA et de la co-expression avec TBX5 dans les 47 tissus et dans le cœur – le lncRNA est fortement exprimé dans le cœur et les deux gènes sont fortement co-exprimés dans les échantillons de cœur.
6) Analyses de l'expression en fonction de l'âge – Les deux gènes sont plus exprimés dans les embryons que dans les adultes.

Cette exploration des données par GEGA suggère que le lncRNA antisens (connu sous le nom de TBX-AS1) pourrait être un régulateur de TBX5 ou du moins pourrait partager une fonction commune. Ce résultat est cohérent avec le fait que le gène du facteur de transcription cardiaque (TBX5) est accompagné d'un lncRNA bidirectionnel comme rapporté en 2018 par Hori et al. [460].

- **Focus sur une région QTL** (*e.g.*, épilepsie de la poule ; Figure 3) : Un QTL pour le caractère épileptique de la poule a été cartographié en 2011 autour des marqueurs 100A3M13 et SEQ1009, qui ont été associés à l'époque au groupe de liaison E26C13 maintenant contenu dans le microchromosome GGA25 [461]. Grâce à une cartographie fine et à des approches moléculaires, les auteurs ont identifié le gène SV2A comme étant lié au caractère épileptique. GEGA aurait pu faciliter l'identification de ce gène de la manière suivante :

1) Définition d'une région de +/- 250kb autour de 100A3M13 et filtration sur le biotype des gènes – 39 PCG sont observés dans la région ;
2) Analyses des termes fonctionnels de ces 39 gènes – 7 gènes répondent à des termes tels que « *brain, neuron, synapse, epilepsy* ».
3) Analyses d'expression pour les 47 tissus – Seul SV2A est spécifiquement exprimé dans le système cérébral, ce qui est cohérent avec le trait d'intérêt.
4) Observation de l'expression des gènes orthologues chez l'humain – Profil d'expression conservé entre les deux espèces.

- **Focus sur une liste de gènes d'intérêt** (e.g., gènes de synthèse et de transport des acides gras ; Figure 4) : Dans le cadre de la réponse adaptative du foie de poule à un changement de source d'énergie alimentaire par le biais de la régulation transcriptionnelle, l'expression de TADA2A apparaissait fortement corrélée à l'expression des gènes liés aux enzymes clés de l'anabolisme des acides gras comme ACACA, FASN, SCD, DLAT, MTTP et ELOVL6 [132]. Grâce à ces résultats, TADA2A a été identifié comme un nouvel acteur potentiel dans la régulation de la lipogenèse. L'analyse de la co-expression de TADA2A avec ces six gènes d'intérêt peut être approfondie en utilisant GEGA et la variété des projets inclus. En utilisant les 11 projets relatifs au foie avec 265 échantillons, la co-expression hépatique de TADA2A avec ACACA, FASN, SCD, LDAT, MTTP et ELOVL6 est confirmée, avec quelques variations ($0,65 \leq r \leq 0,87$) et avec les corrélations les plus élevées pour ACACA et FASN ($r^2 \geq 0,83$). Sachant que les acides gras sont essentiels pour les fonctions cérébrales [462] et qu'une co-expression entre TADA2A

et ACACA dans le cerveau de la souris a déjà été rapportée précédemment [132], la co-expression dans les tissus cérébraux a également été explorée chez la poule. TADA2A est fortement co-exprimé dans l'hypothalamus à l'âge adulte (31 semaines) avec ACACA, FASN, LDAT, ELOVL6 mais pas avec SCD & MTTP. La co-expression dans le cerveau en fonction de l'âge montre que TADA2A est fortement co-exprimé avec ACACA et FASN à l'âge embryonnaire et adulte, il est positivement et négativement co-exprimé à l'âge adulte et embryonnaire pour DLAT et ELOVL6 ; il n'est pas co-exprimé avec SCD et MTTP. En résumé, grâce à GEGA, TADA2A apparaît fortement co-exprimé avec les gènes ACACA & FASN codant les deux enzymes clés de la lipogenèse, quels que soient les projets, les âges et les tissus analysés (foie ou cerveau/hypothalamus) alors que le modèle de co-expression avec DLAT, ELOVL6, SCD et MTTP varie au travers les différentes conditions/tissus.

1.2.3. Discussion et conclusion

L'outil GEGA présenté dans cet article intègre et synthétise les annotations de référence de RefSeq et Ensembl avec celles de projets collaboratifs internationaux tels que FAANG et NONCODE [125]. Bien que chacune des bases de données de référence utilise son propre identifiant (NCBI RefSeq:LOCxx ; Ensembl/GENCODE:ENSGALGxx), elles peuvent toutes deux être utilisées pour travailler avec GEGA en plus du nom HGNC du gène. Cependant, le modèle de gène considéré dans GEGA correspond d'abord à celui de RefSeq et ensuite, s'il n'est pas disponible, à celui d'Ensembl ou des autres bases de données utilisées selon leur ordre d'intégration dans l'annotation enrichie. L'outil GEGA a pour objectif d'améliorer dans les prochaines mises à jour l'annotation des gènes en offrant une annotation plus précise des transcrits composant le modèle génique. En attendant, l'accumulation des six annotations augmente non seulement le nombre de modèles de gènes, mais révèle également la répétabilité de ces derniers et atteste en partie de leur fiabilité.

Au-delà de la simple annotation des loci génétiques, GEGA apporte une réelle valeur ajoutée fonctionnelle en facilitant l'accès aux profils d'expression des gènes au travers de 47 tissus et 1400 échantillons regroupés en 36 jeux de données représentant la diversité des systèmes physiologiques de la poule et via le sexe ou une variété d'âge. Bien qu'elles constituent une base utile, les expressions génétiques de GEGA présentent certaines limites, comme les métadonnées parfois fragmentaires associées à chaque échantillon, qui limitent les

possibilités d'analyse, par exemple pour des conditions spécifiques telles que le stade de développement ou le sexe. Cette limitation peut être surmontée à l'avenir par l'ajout de nouveaux échantillons avec des métadonnées bien définies et qui pourraient facilement être ajoutés de manière cohérente et standardisée en utilisant le *pipeline* « rnaseq » [451] de *nfcore* et l'annotation du génome enrichie fournie. De plus, l'ajout de termes GO, de relations d'orthologie avec la souris et l'homme et d'informations sur la co-expression entre les gènes apporte un éclairage fonctionnel précieux à l'interprétation des données génomiques. Cependant, les bases de données de référence actuelles ne fournissent que des relations d'orthologie entre PCG et quelques miRNA. L'intégration de l'orthologie des lncRNA dans GEGA se fera donc en parallèle avec les avancées dans ce domaine (voir Introduction §1.3 ; Résultats §1.4). Un des enjeux de GEGA est de le maintenir à jour au fur et à mesure de l'évolution des assemblages et des annotations de génomes associées. Le choix a été fait de ne mettre à jour l'annotation que pour les nouvelles versions des assemblages, tout en facilitant le passage de l'une à l'autre et en permettant aux utilisateurs de travailler avec les anciennes versions. Pour un assemblage donné (actuellement GRCg7b), les versions des annotations de référence sont fixes (actuellement RefSeq V106 et Ensembl V107).

En conclusion, malgré ces limitations, GEGA fournit déjà un cadre analytique puissant pour l'exploration fonctionnelle du génome de la poule. Couplé aux dernières avancées de la génomique et de l'annotation des génomes, cet outil bio-informatique soutient la caractérisation fonctionnelle du transcriptome de cette espèce modèle et ouvre la voie à une caractérisation affinée du fonctionnement des génomes complexes.

1.2.4. Valorisation associée

Ces travaux ont fait l'objet :

- d'un article en finalisation d'écriture : **Degalez F**, Bardou P, Lagarrigue S (2023). GEGA (Gallus Enriched Gene Annotation): an online tool gathering genomics and functional information across 47 tissues for 78,323 protein-coding genes and lncRNAs including Ensembl & Refseq genome annotation. Sera soumis à Nucleic Acid Research. **Après son dépôt sur bioRxiv, cet article sera soumis à Nucleic Acid Research. A date de publication de la présente thèse, la dernière version du papier est reproduite ci-après ;**

MANUSCRIPT TITLE

GEGA (Gallus Enriched Gene Annotation): an online tool gathering genomics and functional information across 47 tissues for 78,323 protein-coding genes and lncRNAs including Ensembl & Refseq genome annotation.

AUTHORS

Fabien DEGALEZ[1,†], Philippe Bardou [2,†] and Sandrine Lagarrigue[1,*]

[1] PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France.
[2] SIGENAE, INRAE, 31326 Castanet-Tolosan, France

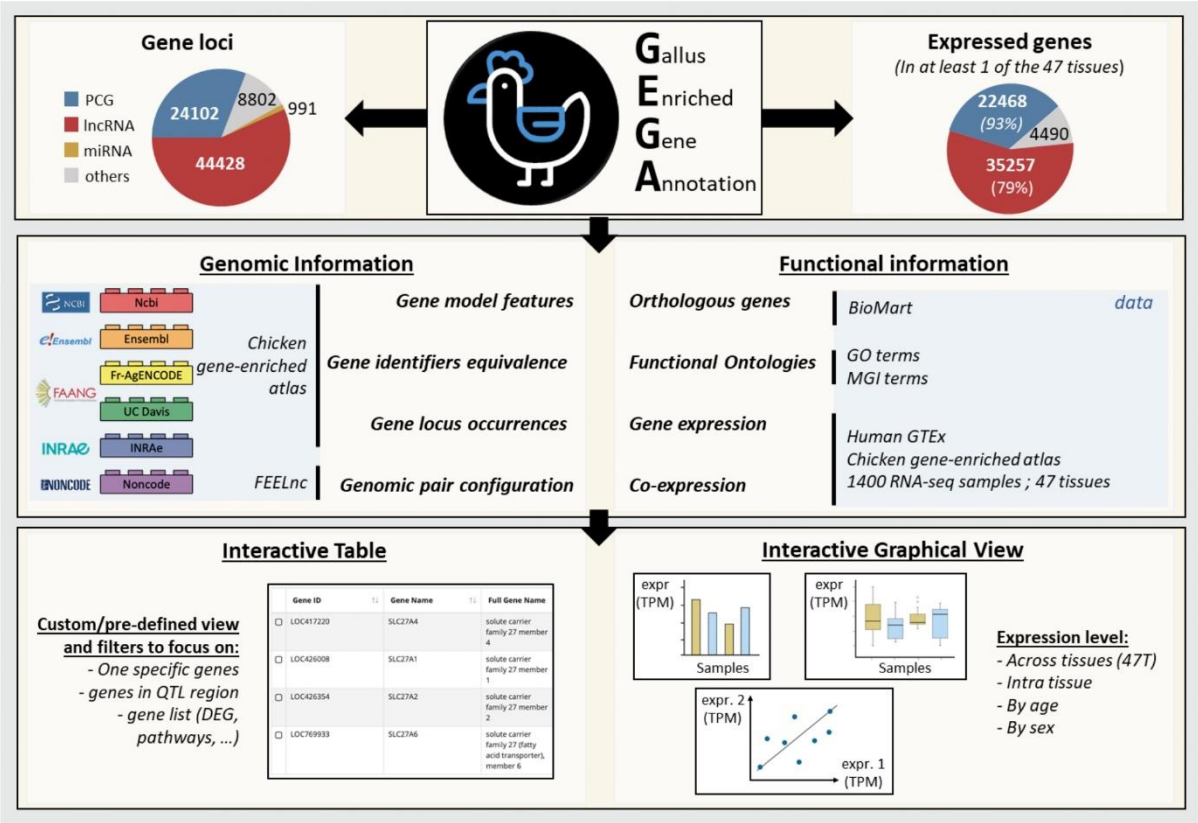
* To whom correspondence should be addressed.

Tel: (+33) 2.23.48.59.59;

Email: sandrine.lagarrigue@institut-agro.fr

† Joint Authors"

13 **GRAPHICAL ABSTRACT**



ABSTRACT

GEGA is a user-friendly tool to navigate through an enriched chicken gene annotation that brings together the power of annotations produced by the reference databases, NCBI-RefSeq & EMBL-Ensembl/GENCODE, and four other additional rich resources as FAANG and NONCODE. With the latest GRCg7b genome assembly, GEGA offers a total of 78,323 genes, including 24,102 protein-coding genes (PCGs) and 44,428 long non-coding RNAs (lncRNAs), greatly enhancing the number of genes provided by each resource separately. But GEGA is more than just a gene locus annotation database. It offers a range of features that allow to go deeper into the functional aspects of these genes by exploring their expression and co-expression profiles across a collection of 47 tissues from 36 datasets and 1400 samples and by discovering tissue-specific variations and study expression as a function of sex or age. For the communities interested in one specific gene, a list of genes or a QTL region, GEGA's user-friendly interface enables efficient gene analysis, easy downloading of results and a multitude of graphical representations, from basic gene information to detailed visualization of expression levels.

INTRODUCTION

The chicken (Gallus gallus) genome is a valuable model in both fundamental and applied research (1). This key model species is often used in order to investigate vertebrate development, evolution and more generally diseases. Moreover, the chicken is one of the livestock species which supplies the most protein-rich food worldwide through both meat and egg production. In recent years, significant progress has been made in annotating this genome, especially thanks to the emergence of high-throughput sequencing technologies and the ongoing efforts of collaborative research projects such as the FAANG (2) and NONCODE (3) consortia or the ChickenGTEx project (4) (part of the FarmGTEx (5)). However, to ensure the most complete annotation of the chicken genome, a lever is to integrate and cross-check annotations from multiple sources using a variety of datasets and analysis pipelines especially the NCBI-RefSeq and EMBL-EBI Ensembl/GENCODE ones which are widely recognized as reputable resources and provide complementary information concerning gene models (6, 7). As previously shown for the last GRCg7b chicken genome assembly, gene annotations between both reference annotations are quite similar in terms of protein coding genes (PCG) with around 90% of the loci shared, even if the transcript models supporting these PCGs are quite different. Concerning long non coding RNA (lncRNA) gene, which are regulatory elements of gene expression contributing to a variety of biological processes, they are harder to identify due to their condition-specific characteristics (tissular & temporal). Consequently, the overlap between their loci reaches around 25% between both annotations (7). In addition to reference gene model databases, the inclusion of additional resources, such as multi-tissue projects or specialized databases, can ease the identification of other gene models. By combining the two reference annotations (and more), the scientific community can exploit the strengths of both databases, providing access to a larger number of gene models and improving the accuracy of their analyses while making easy use of the tools provided by each database. However, even if some initiatives, such as the MANE project for human (8), aims to determine common gene models between these annotations, to date and to our knowledge, an annotation resulting from the union of these reference databases is not currently accessible in an easy and systematic way. The availability of as complete as possible genome annotation is consequently essential for the research community notably in chicken, where genes, especially lncRNAs, are still partially or not known. In this context, we recently released an updated gene atlas of the chicken genome based on the last GRCg7b assembly, integrating reference annotations from both “NCBI-RefSeq” and “EMBL-EBI Ensembl/GENCODE” databases and additional resources as presented in Degalez et al., 2023 (7) . This annotation consists of 78,323 gene models including 24,102 PCGs and 44,428 lncRNAs with a total of 63,513 (81%) genes considered as expressed (≥0.1 TPM) *i.e.*, 22,468 (93%) PCGs and 35,257 (79%) lncRNAs. Indeed, all these genes were functionally annotated

through their expressions across 47 tissues and by analyzing 1400 RNA-seq samples from 36 dataset chosen for well representing the diversity of the chicken's physiological systems. This gene atlas eases switching between "EMBL-EBI Ensembl/GENCODE" and "NCBI-RefSeq" gene identifiers, as well as between assembly versions, since it provides gene identifier equivalents for the previous galgal5 and GRCg6b assemblies of the chicken. This gene loci resource is also complemented by various functional annotations such as orthologous genes in human and mouse, gene ontologies (functional Gene Onthology terms or phenotype terms) or by configuration with the nearest genes

To easily access and explore all this information, we have developed an online tool called GEGA (Gallus Enriched Gene annotation). GEGA is addressed to different communities, whether those interested in genomic regions associated with phenotypes of interest (QTL) or those working directly on gene expression and interested by one specific gene or a list of genes. In this paper, the GEGA tool and the different usages are described.

MATERIAL AND METHODS

Webserver

TODO

Interactive graphical plotting

TODO

Origin and processing of data used in GEGA are detailed in a more exhaustive way in the associated paper, Degalez et al., 2023 (7). Only a summary is provided here.

Reference assembly

The genome annotation is established on the bGalGal1.mat.broiler.GRCg7b (GCF_016699485.2) assembly of the chicken (Gallus Gallus) genome (9).

Data origins – Individual database

The enriched genome annotation included in GEGA integrates gene models from six sources: the "NCBI-RefSeq" (v106) (10) and "EMBL-EBI Ensembl/GENCODE" (v107) (11) reference annotations according to GRCg7b; the "FR-AgENCODE" (12) and "FarmENCODE" FAANG pilot project annotations (13, 14) produced under GRCg6a assembly; the gene annotation from Jehl et al., 2020 (15) under the galgal5 assembly and produced by INRAE; and, finally, the NONCODE annotation (3) including non-coding gene models from galgal4. Note that if FarmENCODE used Oxford Nanopore long-read sequencing, all the other projects mainly used short-read RNA-seq data. Gene models from assemblies older than GRCg7b were remapped using the NCBI genome remapping service with default parameters (16). Quality of gene models was evaluated by the overlap degree of transcripts with FANTOM5 CAGE peaks (17) remapped from galgal5 to GRCg7b. Considering also the database popularity, gene models were added sequentially resulting in this order of aggregation: 1) RefSeq; 2) Ensembl; 3) FrAg; 4) Davis; 5) Inrae; 6) Noncode. Gene models were added from each database to the growing atlas, if their associated transcripts did not overlap genes already present in the growing atlas. Two genes were considered as overlapping if one of their transcripts shared at least one exonic base pair on the same strand using BEDtools (18). To limit overlapping similar patterns with different biotypes, models were aggregated by biotype class.

Data origins – Expression

36 datasets publicly available including a total of 1400 samples were chosen to represent a variety of 47 tissues. The list of the 47 tissues, their abbreviations and colour codes used are available on GEGA.

For each tissue in each project, a median of TPM normalized expressions across samples was calculated. For tissues present in several projects, the median was calculated using the TPM medians previously calculated in each project. A gene was considered as expressed if *i)* its median expression was ≥ 0.1 TPM in at least one tissue, *ii)* at least 50% of samples of a tissue for a given project have a reads number ≥ 6, *iii)* the normalized TPM and TMM expression ≥ 0.1.

Tissue-specificity analysis

Two indicators of tissue specificity are available: *i)* the tau metric (τ), assessed using the log10 of median tissue expression in TPM (19). A gene was considered as tissue specific for a $\tau \geq 0.90$; *ii)* the fold changes between the first and second most expressed tissues.

Classification according to the closest feature

PCG, lncRNA, miRNA and snRNA transcripts and genes were classified through their associated transcripts relatively to their closest PCG and lncRNA transcript using FEELnc v.0.2.1 (default parameters, 100kb between the TSS of the transcripts of the gene of interest and the transcripts of its closest lncRNA or PCG gene model) (20). Briefly, gene pairs are split into two categories and three sub-categories. Firstly, the gene of interest in the pair is considered to be "genic" if it overlaps the partner gene, and "intergenic" otherwise. Secondly, the gene of interest is classified according to its configuration with its partner: "same strand / sense" if it is transcribed in the same orientation; "divergent / antisense" if it is transcribed in head-to-head orientation and; "convergent / antisense" if it is oriented in tail to tail.

Co-expression analysis

For each lncRNA:PCG, lncRNA:lncRNA and PCG:PCG pairs, the Kendall correlation (τ) between the expression values across tissues was computed.

GTEx data analysis

The median gene level TPM for 53 tissues from RNA-seq data of GTEx Analysis V8 was used (https://gtexportal.org/home). The list of the 53 tissues, their abbreviations and color codes used are available on GEGA. In order to compare gene expression, a list of 23 equivalent tissues between humans and chickens has been established and is available on GEGA.

Orthology and GO terms

Gene orthology between the chicken, the mouse and the human as well as GO terms were extracted using BioMart from Ensembl (V107).

RESULTS

General description of the GEGA tool

GEGA is an online tool enabling the exploration of an atlas of 78,323 gene models including both genomics and functional annotations according to the chicken GRCg7b genome assembly. Two related main features are available (Figure 1): *(i)* an interactive table with possibilities to view different gene features as presented briefly in the "Mat. and Met."; *(ii)* an interactive viewer providing plots to explore the different patterns of expression and co-expression for gene models according to different conditions (inter tissues / intra tissue; sex / age) but also to explore gene models information previously selected. This tool is publicly available on https://gega.sigenae.org/ and free to use. It does not depend on cookie files or any credentials. The web server was tested on Chrome V.116.0, Firefox V.116.0 and Safari V16.5.2. browsers. A dedicated contact tab is available to help improve the tool, particularly if ideas for improvement, conception problems or bugs are identified.

Interactive Table – Data viewing

A variety of data are available, but the user may not be interested in all of it. Therefore, only data concerning the gene model *strincto sensu* are initially displayed by default. The user can display additional data through the dedicated "Views" panel (Figure 1A). Pre-filtered views were implemented to guide potential new users or facilitate access to specific information, following the main categories of information available in GEGA. The first three columns, which are always identical, include the gene identifier from the source and the short and long gene names. The preconfigured views are as follows: *(i)* "Default", displaying the origin, positional information, and biotype; *(ii)* "Naming", showing available identifiers for a gene model, particularly Ensembl and NCBI, and for previous assemblies including galgal5 and, GRCg6a; *(iii)* "Expression & Tissue-specificity", indicating expression values (in TPM) in each of 47 tissues, top tissues, and tissue specificity; *(iv)* "Functional terms", listing gene ontology (GO) and phenotype terms for each gene and their associated orthologs; *(v)* "Orthology", showing orthologous relationships and equivalents with human and mouse for each protein coding gene; *(vi)* "Gene structure", detailing transcript, exon, and intron structure; *(vii)* "FEELnc" indicating the nearest gene model (PCG/lncRNA/miRNA) and configuration at the gene and transcript level, along with co-expression data at the gene level; *(viii)* "Occurrence", providing reproducibility information on loci representation across resources. For flexibility and specific usage, columns can be manually selected with the "Customize" tool using names or the tree structure with checkboxes.

Interactive Table – Data filtering

Filters can be easily applied to extract a specific set of genes using the "Search" panel (Figure 1B). All columns available in the "Views" panel can be filtered and combined. Some filters can be applied simultaneously using predefined "custom function" available in the same selection menu. To date, the

"Function Gene name global search" and "Function Functional term search" allow to filter across all the naming (names and gene identifiers from both Refseq and Ensembl) or GO related columns respectively. Logical operators can be applied according to the data type. The "equals/not equals" operator can be used for both numerical and categorical data, while "greater/lower" is specific to numerical data and "contained/not contained" to categorical data. For a given filter, the comma (",") delimiter can be used to indicate several possibilities (OR operator), *e.g.,* application of the filter "gnSimpleBiotype – contains – lnc,pcg" will results of a selection of both lncRNA and PCG. In the same way, "Function Gene name global search – contains – LOC417220, ENSGALG00010025549, ACACA" will result in a selection of the three genes mentioned, even if the type of nomination used varies. Note that, each line corresponds to a criterion and a set of criteria (AND operator) is considered as an intersection of condition. Even if each tissue expression is not available individually through the interactive table, filter on tissue gene expression is available through the "[tissueName]_expr" filter. Note that filters apply even if the column is not displayed, *e.g.*, a genomic region can be selected while only displaying the functional view.

In order to answer to specific usage (see user cases), three modules were designed to easily generate the filters required for the intended use: *i)* "Gene list" module has been implemented to quickly input a gene list and view the associated data. *ii)* "Define a region" module creates a region surrounding a gene (using identifier or name) or position (defined by chromosome and position), which can be used, *e.g.,* for QTL analysis. Offsets can be applied equally or differently on each side. *iii)* "Expression threshold" module enables to easily apply a combination of filters based on expression criteria.

Interactive Table – Output data

After filtering, genes of interest (*i.e,.* lines of the table – Figure 1C) can be easily selected and either gene names, gene identifiers or the genomic region can be copied to the clipboard (Figure 1D). The selected list of genes or the genomic region can be then paste elsewhere for further analysis or for an interactive visualisation.

Two types of filtered tables are available for download (Figure 1E) as either a *.tsv* or *.xlsx* file depending on user preference: *(i)* "Annotation", which includes all available information or only what is presented in the custom view as chosen by the user; "Expression" which provides individual expression data for the filtered genes in the 47 tissues.

Interactive visualization – Explore genes according to the filtered interactive table

An interactive visualizer provides a graphical overview of gene sets selected by specific filters (Figure 1F). Consequently, pie charts and boxplots can be generated for categorical and numerical variables,

respectively. Considering two numerical variables, a scatter plot showing potential relationships between them can be plotted, *e.g.*, lncRNA and PCG co-expression *vs.* separating distance.

Interactive visualization – Global usage to explore the gene expression

Due to data variability, expression can be observed at different levels. While expression patterns can be analyzed by comparing across 47 tissues, variation between individuals and projects (when available) can also be observed within a specific tissue. Similarly, after selecting a tissue of interest, a project (or all available) can be chosen to view gene expression by criteria such as sex or ages. Although each feature has its own specificity, boxplots and co-expression plots (scatter plots) are always available and work similarly. For boxplots, a gene or list of genes (separated by spaces or commas) can be provided. Expression for each gene will be displayed in the input order. To enable comparison, expression can be unified by displaying all values on a scale of 0 to the maximum observed across all genes. Sample identities (individuals) can also be displayed. Tissues are classified by abbreviation by default, but can be ordered by expression proximity obtained by classification (see Degalez et al., 2023 (7)). Tissues sharing color ranges generally share a physiological system. For co-expression plots, a reference gene identifier is first provided, followed by a target gene(s). Each plot of the reference (X-axis) versus target (Y-axis) gene is displayed, initially in input order. However, plots can be ordered by correlation (1 to -1) or filtered by a correlation threshold. Because co-expression can follow different laws, linear or logarithmic plots are available. Displaying the regression is also an option.

Interactive visualization – Specific usage: "Expression intra tissues" (Figure 1G)

First, after selecting the tissue of interest, expression profiles across 47 tissues can be displayed for the 40 most highly expressed genes. This enables identification of tissue-specific or ubiquitous genes. Finally, after choosing a tissue of interest, boxplots of expression by project or co-expression plots integrating all available samples can be displayed.

Interactive visualization – Specific usage: "Expression across tissues" (Figure 1H)

First, to overview gene expression correlation across 47 tissues, a correlation plot is available. After a gene identifier is provided, the 50 most correlated genes are displayed in descending correlation order. The gene list is accessible and can be copied/pasted for further investigation. For visualization, genes on chromosomes other than the one requested are shown in grey; genes on the same chromosome appear in green if their distance is less than the threshold, otherwise blue. This can help to identify clustered co-expressed genes, regardless or not of location. As an example, TBX5 (LOC373988; chr15:12,317,331) shows a high co-expression across tissues (ρ=0.84) with a lncRNA (ENSGALG00010000570; chr15:12,314,252) which is in 15th position and coloured in green because both genes are separated by less than 100kb (default value).

For examining expression across tissues, after providing a gene or a gene list, either a boxplot (as described before) or barplot of expression across the 47 tissues can be generated. Boxplots consider all samples from all projects together, plotting all available data. Barplots first calculate the median expression for each project, then the median of those medians, thus lacking sample resolution. Barplots have the same options as boxplots.

Interactive visualization – Specific usage: "Expression by age/sex" (Figure 1I)

Because each project has its own experimental design, especially regarding sex and age differences, a project of interest must first be selected. This ensures reliable and consistent sub-group comparisons across studied factors. All projects can be analyzed together, particularly to increase condition or sample numbers, but result interpretation is left to user discretion. After selecting the project and tissue of interest, barplots of gene expression by condition and sample can be generated on linear or logarithmic scales, with unified or raw maximum values. Boxplots by condition can be generated for one specific tissue or up to four tissues. In addition to previously described options, boxplots can be ordered by condition or tissue depending on what the user wants to highlight. Finally, co-expression plots using all available samples for a given tissue can be displayed, following the previously described process.

Interactive visualization – Specific usage: Orthology with the GTEx.

TODO

Interactive visualization – Output data

Regardless of plot type, the associated figure can be downloaded in PNG, JPEG, SVG, or PDF format through the dedicated menu (top-right of plot). The data used for plotting can be downloaded as a *.tsv* or *.xlsx* file, formatted to enable re-plotting with specific tools (e.g. R-base, ggplot) as desired by the user. Following this principle, production of multi-plots is currently unavailable; each plot and associated data must be downloaded independently.

GEGA usages through typical user cases

- ***Focus on one specific gene (e.g., TBX5; Figure 2)***

TBX5, a T-box transcription factor (21), plays a critical role in vertebrate cardiac morphogenesis. Overexpression of TBX5 in embryonic chick hearts has shown that TBX5 inhibits myocardial growth cardiogenesis and thereby participate in modulating vertebrate cardiac growth and development (22). Using GEGA, this gene can then be deeply explored, *e.g.*, following this process: 1) Analysis of expression across the 47 tissues and 2) through functional terms (GO terms / phenotype terms) to underline those which are consistent with the TBX5 function. 3) Observation of genomics configuration between the two neighboring genes. This feature underlines the presence of an antisense lncRNA. 4) Gene orthology examination for human and mouse. The TBX5 gene has an

orthologous PCG in mouse and human. This initial observation served manual verification of the conservation in both species of the antisense lncRNA. 5) Analyses of expression and co-expression across the 47 tissues. TBX5 and its antisense lncRNA revealed that the lncRNA is highly expressed in heart as TBX5. Moreover, both genes are highly co-expressed in heart samples. 6) Analyses of expression related to ages. The gene pair is more expressed in embryo than adult stages.

This data exploration by GEGA suggests that the antisense lncRNA (known as TBX-AS1) could be a regulator of TBX5 or at least could share a common function. This result is consistent with the fact the cardiac transcription factor gene (TBX5) is accompanied by a bidirectional long non-coding RNA as reported in 2018 by Hori et al. (23).

- ***Focus on a QTL region (e.g., chicken epilepsy; Figure 3)***

A QTL for epilepsy trait in chicken was mapped in 2011 around the markers 100A3M13 and SEQ1009, that mapped at that time to linkage group E26C13 which has been identified thereafter as the microchromosome GGA25 (24). Through fine mapping and other molecular approaches, the authors identified the SV2A gene as related to the epilepsy trait. GEGA could have facilitated the identification of this gene as follows: 1) Definition of a region of +/- 250kb around 100A3M13 and display of all the genes, with the possibility of filtering the biotype of the genes (protein-coding genes (PCG), long non-coding genes (lncRNA, etc). In our case, 39 protein-coding genes are observed in the region. 2) Analyses of the functional terms of these 39 genes. Of these, 7 genes respond to GO terms or phenotypes associated with the trait of interest, *i.e.,* "brain, neuron, synapse, epilepsy". 3) Expression analyses on the 47 tissues. Only SV2A appears to be specifically expressed in the cerebral system, which is consistent with the trait of interest. 4) Observation the orthologous gene expression in human shows that the expression pattern is conserved between both species.

- ***Focus on a gene list (e.g., fatty acid synthesis and transport genes; Figure 4)***

The main role of liver in the chicken adaptive response to a switch in dietary energy source through the transcriptional regulation of lipogenesis was previously reported. Indeed, 298 down-regulated genes in "High Fat / High Fiber diet (HF)" compared to a standard diet containing "Low Fat / High Starch" (LF) with an enrichment of genes related to fatty acid synthesis and transport was observed (25). The expression in liver of TADA2A (Transcriptional Adaptor 2A) was highly correlated to gene expression related to key enzymes of fatty acids (FA) anabolism as ACACA, FASN, SCD, DLAT, MTTP and ELOVL6. Thanks to these results, TADA2A has been identified as a potential new player in the regulation of lipogenesis. Analysis of the co-expression of TADA2A with these 6 genes of interest can be further investigated using GEGA and the variety of projects included. Using the 11 projects related to liver with 265 samples, the hepatic co-expression of TADA2A with ACACA, FASN, SCD, LDAT,

303 MTTP and ELOVL6 is confirmed, with some variations ($0.65 \leq r \leq 0.87$) and with the highest correlations
304 for ACACA and FASN ($\rho \geq 0.83$). The co-expression in brain was also explored. Indeed, fatty acids are
305 essential for brain functions (26) and a co-expression between TADA2A and ACACA in mouse brain
306 was already previously reported (25). In the liver, TADA2A is highly co-expressed in hypothalamus at
307 an adult age (31 week of age) with ACACA, FASN, LDAT, ELOVL6 but not with SCD & MTTP. Co-
308 expression in the brain as a function of age shows that TADA2A is highly co-expressed with ACACA and
309 FASN across embryo and adult ages, it is positively and negatively co-expressed across adult and
310 embryo ages for DLAT and ELOVL6; it is not co-expressed with SCD and MTTP. In summary, though
311 GEGA analysis, TADA2A appears highly co-expressed with ACACA & FASN genes coding the two key
312 enzymes of lipogenesis, whatever the projects, the ages and the analyzed tissues (liver or
313 brain/hypothalamus) whereas the co-expression pattern is multi-form across these different
314 conditions/tissues with DLAT, ELOVL6, SCD and MTTP.

DISCUSSION

The GEGA tool presented in this article integrates and synthesizes the reference annotations from NCBI-RefSeq and Ensembl/GENCODE with those from international collaborative projects such as FAANG and NONCODE (7). Although each of the reference databases uses its own identifier (NCBI-RefSeq:LOC$_{xx}$; Ensembl/GENCODE:ENSGALG$_{xx}$), both can be used working with GEGA in addition to the HGNC name of the gene. However, the gene identifiers used may not correspond strictly to the associated gene model. Indeed, the gene model considered in GEGA corresponds firstly to that of NCBI-RefSeq and then, if not available, to that of Ensembl/GENCODE ones. The GEGA tool will be improved in future updates in parallel with the annotation of the associated gene, which may offer an annotation at transcript level. The accumulation of the six annotations not only increases the number of gene models but also reveals the repeatability of these models through these annotations. Even though it can be difficult to assess the accuracy of the gene models, particularly in terms of exon/intron structure or gene fusions/splits, due to variable samples and different analysis methods, the repeated presence of a model at a specific locus in various annotations supports the existence of a gene model, especially for lncRNAs. Beyond the simple annotation of gene loci, GEGA provides real functional added value by integrating the expression profiles of these genes across 47 tissues and 1,400 samples grouped into 36 datasets representing the diversity of chicken physiological systems. As a result, 63,513 genes out of the 78,323 annotated (*i.e.,* 81%) are considered to be expressed. This standardized functional annotation greatly facilitates transcriptomic studies in this species by providing a reliable expression reference. Moreover, GEGA's interactive graphical interface makes it easy to explore this expression data according to different physiological conditions (tissue, sex, age), paving the way for new discoveries. Although providing a useful basis, the gene expressions in GEGA have certain limitations such as the sometimes fragmentary metadata associated with each sample, which limits the possibilities of analysis, for example for specific conditions such as stage of development or sex. This limitation may be overcome in the future by adding new projects (which can be suggested through the “Contact” tab), with metadata well defined, which could easily be added in a coherent and standardized manner using the nfcore “rnaseq” pipeline and the genome annotation provided enriched with gene models. Furthermore, the addition of GO terms, orthology relationships with mice and humans and co-expression information between genes sheds valuable functional light on the interpretation of genomic data. The chicken is a model species for many human diseases, and comparisons of expression profiles between species provide a wealth of information. Similarly, analysis of co-expression networks between coding and non-coding genes highlights potential transcriptional regulations and opens up promising research prospects. However, current reference databases only provide orthology relationships between PCGs and a few miRNAs. The integration of lncRNA orthology into GEGA will therefore be carried out in parallel with advances in this field. One of

the issues with GEGA is maintaining it up to date as genome annotation evolves, in particular with the aim of producing complete genomes from telomere to telomere (T2T) (27). The choice was made to update the annotation only for new versions of genome assemblies, while facilitating the transition from one to the other and enabling users to work with older versions. For a given assembly (currently GRCg7b), the versions of the reference annotations are fixed (currently NCBI-RefSeq V106 and Ensembl/GENCODE V107). In conclusion, despite these limitations, GEGA already provides a powerful analytical framework for functional exploration of the chicken genome. Coupled with the latest advances in genomics and genome annotation, this bioinformatics tool supports the functional characterization of the transcriptome of this model species and paves the way for a refined characterization of the functioning of complex genomes.

REFERENCES

1. Burt,D.W. (2007) Emergence of the chicken as a model organism: implications for agriculture and biology. *Poult Sci*, **86**, 1460–1471.

2. Giuffra,E., Tuggle,C.K., and FAANG Consortium (2019) Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu Rev Anim Biosci*, **7**, 65–88.

3. Zhao,L., Wang,J., Li,Y., Song,T., Wu,Y., Fang,S., Bu,D., Li,H., Sun,L., Pei,D., *et al.* (2021) NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res*, **49**, D165–D171.

4. Guan,D., Bai,Z., Zhu,X., Zhong,C., Hou,Y., Consortium,T.C., Lan,F., Diao,S., Yao,Y., Zhao,B., *et al.* (2023) The ChickenGTEx pilot analysis: a reference of regulatory variants across 28 chicken tissues. 10.1101/2023.06.27.546670.

5. FarmGTEx - Farm Animal Genotype-Tissue Expression FarmGTEx.

6. Lagarrigue,S., Lorthiois,M., Degalez,F., Gilot,D. and Derrien,T. (2022) LncRNAs in domesticated animals: from dog to livestock species. *Mamm Genome*, **33**, 248–270.

7. Degalez,F., Charles,M., Foissac,S., Zhou,H., Guan,D., Fang,L., Klopp,C., Allain,C., Lagoutte,L., Lecerf,F., *et al.* (2023) Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues. 10.1101/2023.08.18.553750.

8. Morales,J., Pujar,S., Loveland,J.E., Astashyn,A., Bennett,R., Berry,A., Cox,E., Davidson,C., Ermolaeva,O., Farrell,C.M., *et al.* (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, **604**, 310–315.

9. NCBI-RefSeq (2021) bGalGal1.mat.broiler.GRCg7b - Genome - Assembly - NCBI. *NCBI*.

10. NCBI-RefSeq (2022) bGalGal1.mat.broiler.GRCg7b - Genome - Annotation - NCBI v106.

11. EMBL EBI Ensembl/GENCODE (2022) bGalGal1.mat.broiler.GRCg7b - Genome - Annotation - Ensembl v107.

12. Foissac,S., Djebali,S., Munyard,K., Vialaneix,N., Rau,A., Muret,K., Esquerré,D., Zytnicki,M., Derrien,T., Bardou,P., *et al.* (2019) Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biology*, **17**, 108.

13. Andersson,L., Archibald,A.L., Bottema,C.D., Brauning,R., Burgess,S.C., Burt,D.W., Casas,E., Cheng,H.H., Clarke,L., Couldrey,C., *et al.* (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology*, **16**, 57.

14. Tixier-Boichard,M., Fabre,S., Dhorne-Pollet,S., Goubil,A., Acloque,H., Vincent-Naulleau,S., Ross,P., Wang,Y., Chanthavixay,G., Cheng,H., *et al.* (2021) Tissue Resources for the Functional Annotation of Animal Genomes. *Frontiers in Genetics*, **12**.

15. Jehl,F., Muret,K., Bernard,M., Boutin,M., Lagoutte,L., Désert,C., Dehais,P., Esquerré,D., Acloque,H., Giuffra,E., *et al.* (2020) An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci Rep*, **10**, 20457.

16. NCBI-RefSeq (2022) Coordinate remapping service: NCBI.

17. Lizio,M., Deviatiiarov,R., Nagai,H., Galan,L., Arner,E., Itoh,M., Lassmann,T., Kasukawa,T., Hasegawa,A., Ros,M.A., *et al.* (2017) Systematic analysis of transcription start sites in avian development. *PLOS Biology*, **15**, e2002887.

18. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

19. Yanai,I., Benjamin,H., Shmoish,M., Chalifa-Caspi,V., Shklar,M., Ophir,R., Bar-Even,A., Horn-Saban,S., Safran,M., Domany,E., *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.

20. Wucher,V., Legeai,F., Hédan,B., Rizk,G., Lagoutte,L., Leeb,T., Jagannathan,V., Cadieu,E., David,A., Lohi,H., *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res*, **45**, e57.

21. Chapman,D.L., Garvey,N., Hancock,S., Alexiou,M., Agulnik,S.I., Gibson-Brown,J.J., Cebra-Thomas,J., Bollag,R.J., Silver,L.M. and Papaioannou,V.E. (1996) Expression of the T-box family genes, Tbx1-Tbx5, during early mouse development. *Dev Dyn*, **206**, 379–390.

22. Hatcher,C., Kim,M., Maha,C., Goldstein,M., Wong,B., Mikawa,T. and Basson,C. (2001) TBX5 transcription factor regulates cell proliferation during cardiogenesis. *Developmental biology*, **230**, 177–188.

23. Hori,Y., Tanimoto,Y., Takahashi,S., Furukawa,T., Koshiba-Takeuchi,K. and Takeuchi,J.K. (2018) Important cardiac transcription factor genes are accompanied by bidirectional long non-coding RNAs. *BMC Genomics*, **19**, 967.

24. Douaud,M., Feve,K., Pituello,F., Gourichon,D., Boitard,S., Leguern,E., Coquerelle,G., Vieaud,A., Batini,C., Naquet,R., *et al.* (2011) Epilepsy Caused by an Abnormal Alternative Splicing with Dosage Effect of the SV2A Gene in a Chicken Model. *PLoS One*, **6**, e26932.

25. Desert,C., Baéza,E., Aite,M., Boutin,M., Le Cam,A., Montfort,J., Houee-Bigot,M., Blum,Y., Roux,P.F., Hennequet-Antier,C., *et al.* (2018) Multi-tissue transcriptomic study reveals the main role of liver in the chicken adaptive response to a switch in dietary energy source through the transcriptional regulation of lipogenesis. *BMC Genomics*, **19**, 187.

26. Hamilton,J.A., Hillard,C.J., Spector,A.A. and Watkins,P.A. (2007) Brain uptake and utilization of fatty acids, lipids and lipoproteins: application to neurological disorders. *J Mol Neurosci*, **33**, 2–11.

27. Huang,Z., Xu,Z., Bai,H., Huang,Y., Kang,N., Ding,X., Liu,J., Luo,H., Yang,C., Chen,W., *et al.* (2023) Evolutionary analysis of a complete chicken genome. *Proceedings of the National Academy of Sciences*, **120**, e2216641120.

DATA AVAILABILITY

The data underlying this article are available through the GEGA website at https://gega.sigenae.org/. The raw data are also available through the FR-AgENCODE website at https://www.fragencode.org/lnchickenatlas.html.

SUPPLEMENTARY DATA

No supplementary data is provided.

AUTHOR CONTRIBUTIONS

No author contributions is provided.

ACKNOWLEDGEMENTS

No acknowledgement is provided.

FUNDING

This project is funded by the European Union's Horizon 2020 research and innovation program under grant agreement N°101000236 (GEroNIMO) and by ANR CE20 under 'EFFICACE' program. Fabien Degalez is a Ph.D. student supported by the Brittany region (France) and the INRAE (Animal Genetics Division).

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare that are relevant to the content of this article.

TABLE AND FIGURES LEGENDS

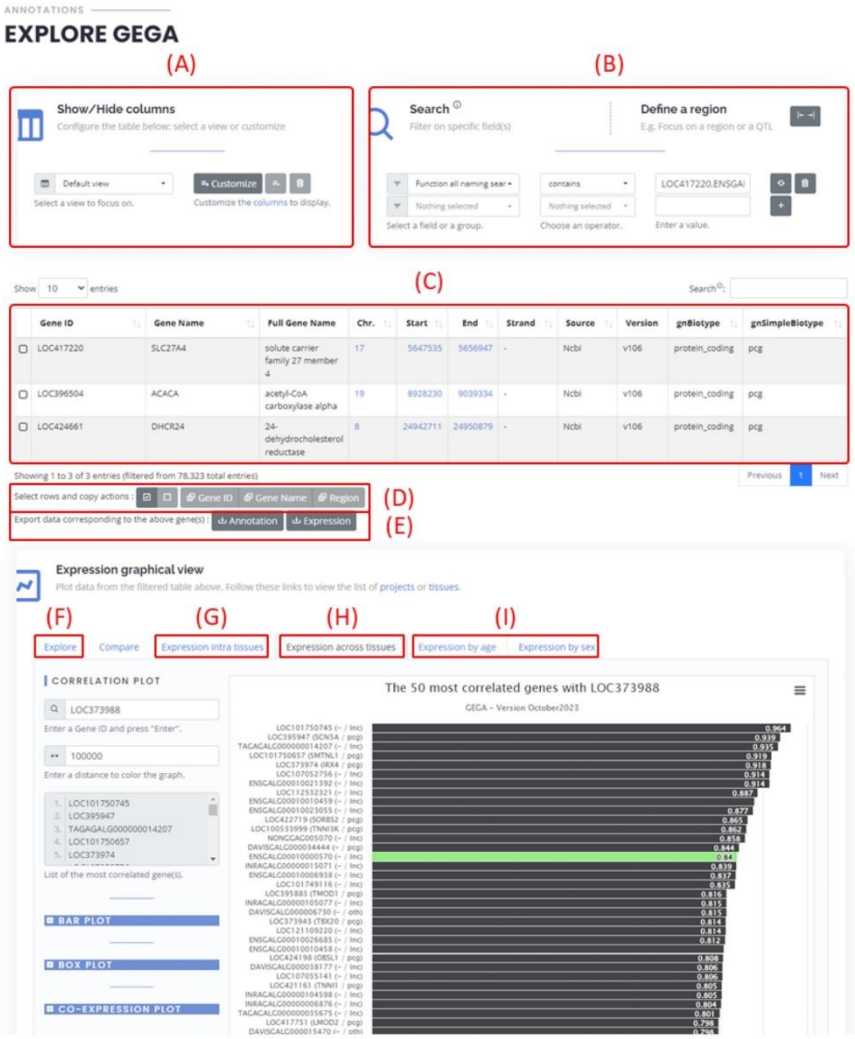
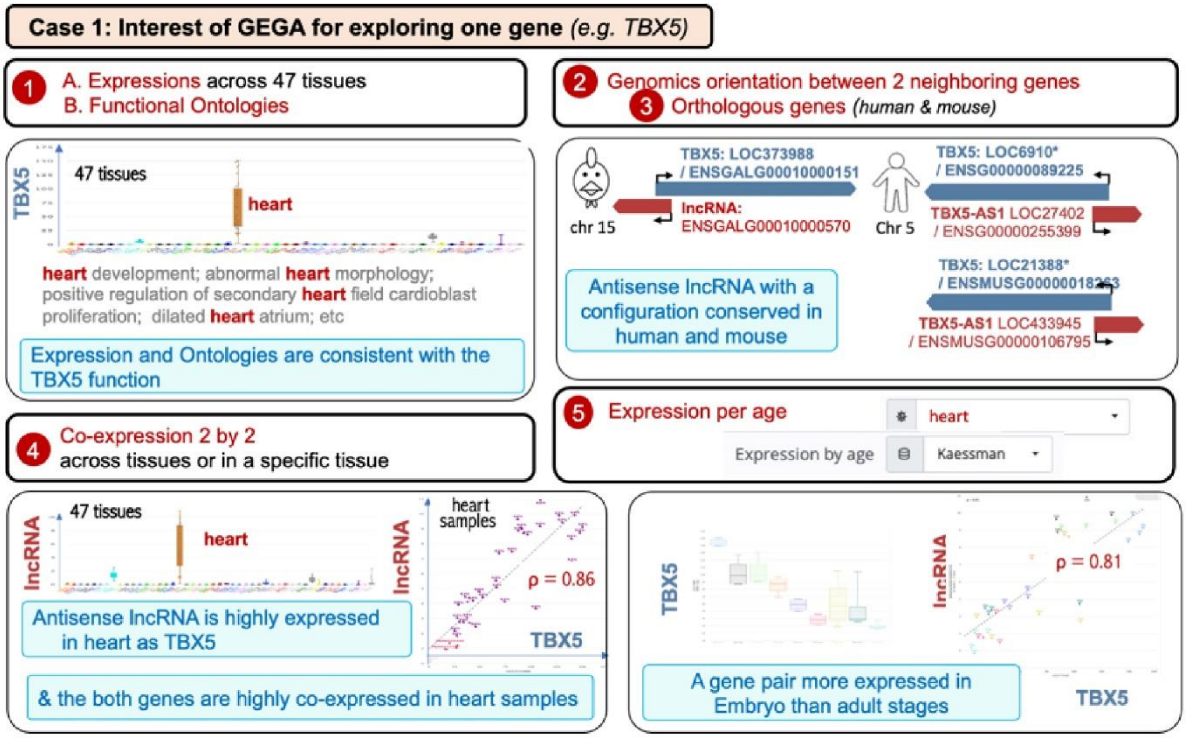
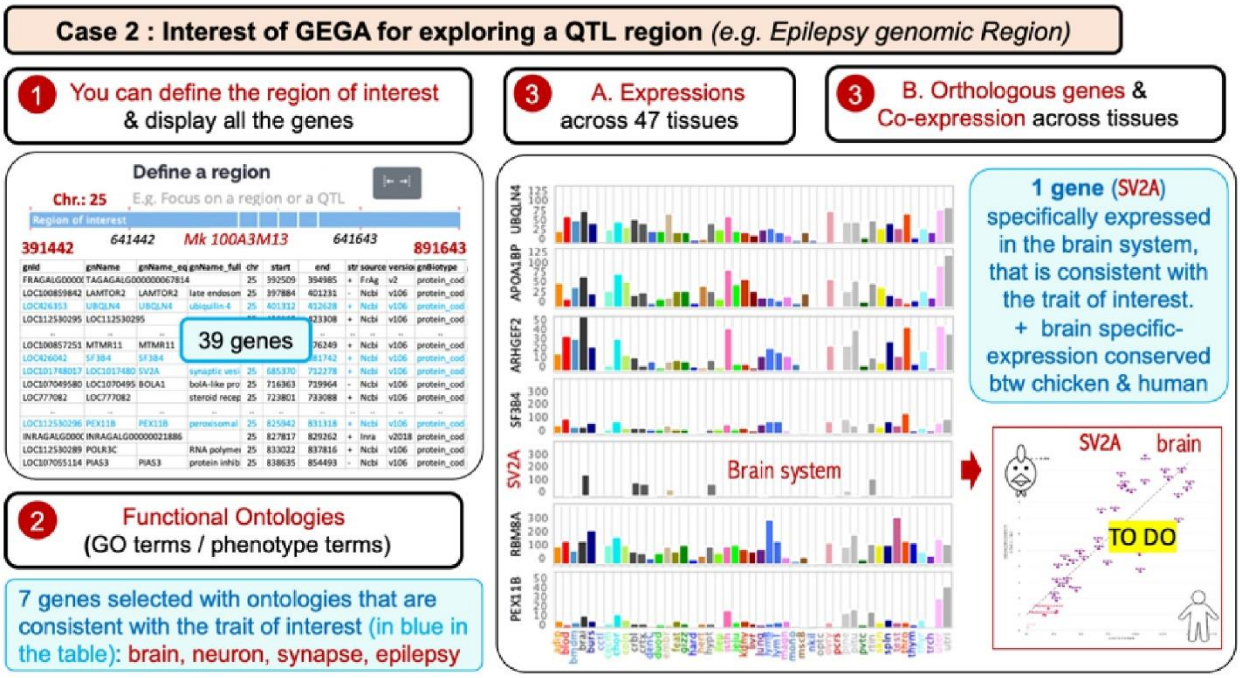


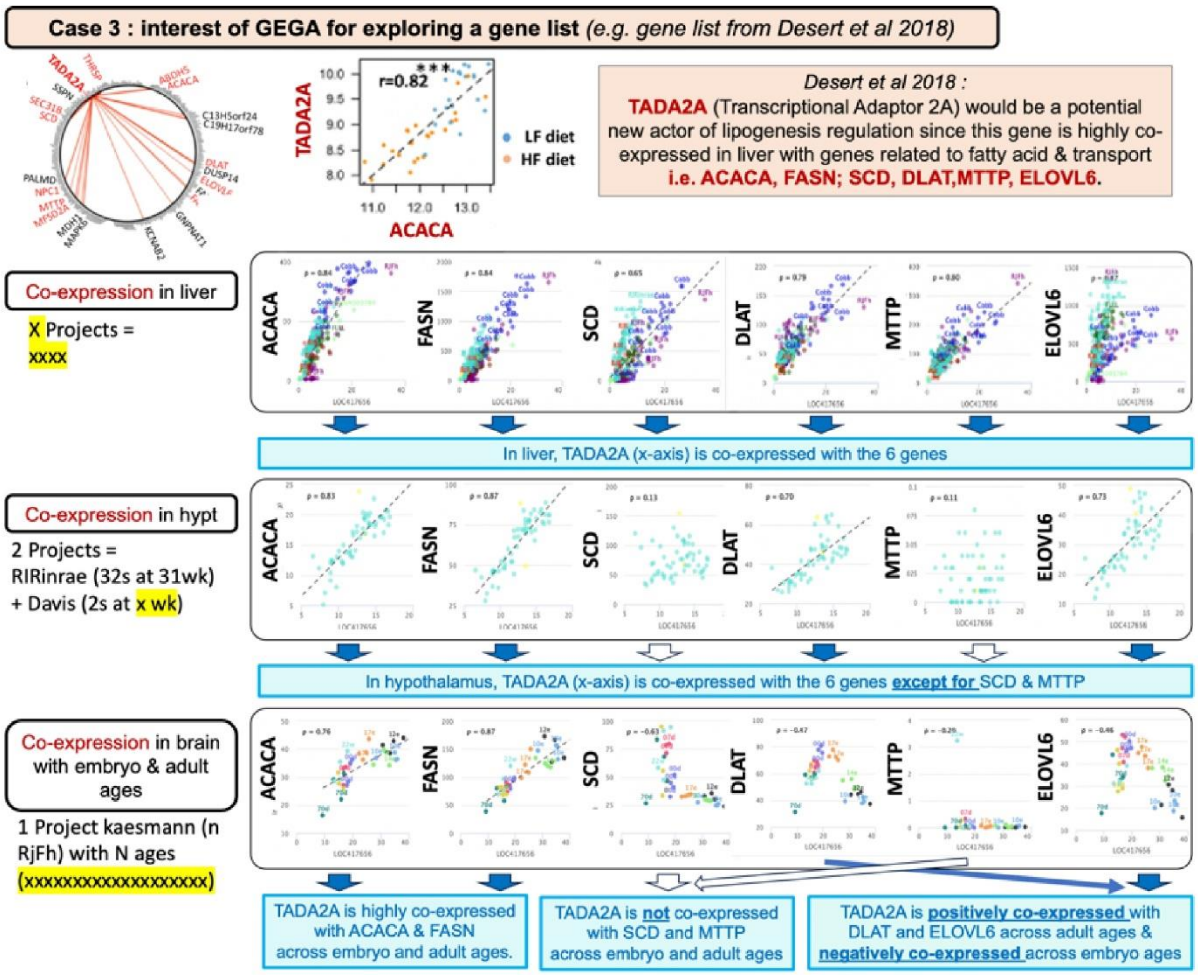
Figure 1. Global application interface. (A) View selection panel. (B) Filter selection panel. (C) Table of selected gene according to the chosen view. (D) Copy panel. (E) Exportation of annotation and/or expression. (F) Visual exploration of the selected genes. Graphical tab for (G) intra-tissue, (H) inter-tissues and (I) age/sex expressions



458 **Figure 2.** User case 1 – GEGA for exploring a specific gene, its genomic environment and its potential
459 functions (*e.g.*, TBX5).



460 **Figure 3**. User case 2 – GEGA for analyzing a QTL region (*e.g*, epilepsy region around the marker
461 100A3M13)



462 **Figure 4.** User case 3 – GEGA for exploring a gene list (*e.g.*, fatty acid synthesis and transport genes)

1.3. Impact de l'assemblage du génome pour l'interprétation des données RNAseq et le diagnostic des maladies rares chez l'humain

(Résumé d'article)

<u>Aparté :</u> En juillet 2022 et pour une durée de trois mois, j'ai eu l'opportunité d'intégrer en tant qu'étudiant-chercheur invité le laboratoire du Dr. Stephen Montgomery (Université Stanford, Californie, États-Unis), un des leaders du consortium GTEx humain. Durant cette période, j'ai pu, entre autres, intégrer un groupe de travail s'intéressant à l'impact de l'assemblage du génome sur l'interprétation du RNAseq et le diagnostic des maladies rares. Pour ma part, j'ai majoritairement travaillé sur les différences d'annotations entre hg19, hg38 et CHM13 pour lesquelles j'ai pu fournir une table d'équivalence et un rapport détaillé des changements observés entre assemblage. Par souci de confidentialité et attendant la parution publique de l'article, les paragraphes qui suivent ne présentent qu'une partie des résultats et se focalisent davantage sur les éléments auxquelles j'ai pu activement participer.

1.3.1. Contexte et objectifs

La transcriptomique est de plus en plus utilisée comme stratégie complémentaire pour le diagnostic des maladies rares [463] . La sélection d'un assemblage de référence et d'une annotation correspondante constituent la base de l'analyse du transcriptome, et tandis que l'impact des différentes sources d'annotation sur les estimations de l'expression génique est bien documenté [464–470], l'impact de l'assemblage est moins bien compris. Malgré la publication des assemblages du génome humain, hg19 en 2009 [471] et de sa version améliorée hg38 en 2013 [472], la plupart des laboratoires universitaires et commerciaux continuent d'utiliser hg19 [473]. Cependant, la récente parution du premier assemblage complet du génome humain – CHM13 du consortium « *Telomere2Telomere* » (T2T) – offre une option supplémentaire dans le choix de l'assemblage et augmente l'incertitude concernant l'impact de ce dernier sur l'analyse du transcriptome pour le diagnostic [474]. Pour évaluer l'impact du choix de l'assemblage sur la quantification générale de l'expression des gènes, mais également dans le cadre de la détection d'expressions aberrantes pour l'étude des maladies rares, une évaluation complète de l'impact du choix des assemblages hg19, hg38 et CHM13 sur les résultats RNAseq avec une résolution au niveau des gènes a été menée. Il a été mis en évidence des cas de maladies rares et non diagnostiquées pour lesquelles le choix de l'assemblage du génome peut avoir une incidence sur le diagnostic. Pour ce faire, les données transcriptomiques d'une cohorte de patients atteints de maladies rares avec des troubles hétérogènes [475] ont été alignées sur les assemblages hg19, hg38 et CHM13. Ainsi, des gènes

i) dont l'expression est spécifique à la présence du gène dans l'annotation, *ii)* différentiellement exprimés entre assemblages ou *iii)* exclusivement exprimés dans l'un d'entre eux, ont été identifiés à l'aide six tissus cliniquement accessibles. Finalement, une ressource indiquant les effets du choix de l'assemblage pour 1 332 gènes est fournie afin de permettre la prise de décision concernant la construction du génome pour les analyses actuelles et futures.

1.3.2. Résultats préliminaires

Cartographie du transcriptome d'une cohorte de maladies rares à travers les assemblages de génome

Les données RNAseq ont été produites à partir de 386 échantillons provenant de 316 personnes. Six tissus étaient représentés dont le sang, les fibroblastes, les cellules mononucléaires du sang périphérique (PBMC), le muscle squelettique, les cellules souches pluripotentes induites (iPSC) et les cellules progénitrices neurales dérivées (NPC). Notons que la majorité des échantillons provenait du sang (n = 283). Cette cohorte comprenait 204 cas présentant principalement des symptômes neurologiques, musculo-squelettiques ou liés au système immunitaire, offrant ainsi une représentation hétérogène des phénotypes de maladies rares. Chaque échantillon a été aligné sur les génomes hg19.p13, hg38.p13 et CHM13v2 de manière uniforme à l'aide d'un *pipeline* standardisé. Pour garantir une cohérence maximale des annotations génétiques, les gènes ont été quantifiés à l'aide des annotations génétiques équivalentes de GENCODEv35 pour chaque construction. En effet, cette annotation présente l'avantage d'avoir été définie sur hg38 puis transférée par *remapping* sur hg19 et CHM13. Soulignons qu'une proportion plus faible de *reads* multi-cartographiées et une proportion plus élevée de *reads* non cartographiés ont été observées pour l'alignement sur CHM13. Ce résultat peut s'expliquer en partie par la proportion accrue de *reads* non cartographiés qui sont classés comme trop courts pour l'alignement, probablement en raison de la complexité accrue de l'assemblage CHM13.

Identification des changements dans l'annotation des gènes entre les assemblages

La cohérence des trois annotations utilisées, toutes dérivées de GENCODEv35, a été étudiée en comparant la structure des exons et des transcrits et la séquence génétique sous-jacente pour chaque gène. Les gènes ont été définis comme ayant des modèles identiques si le nombre de transcrits et d'exons constitutifs et la longueur des exons étaient identiques entre les modèles. Pour les gènes ayant des modèles identiques, la similarité des séquences a été calculée à l'aide du score de similarité de Jaro-Winkler [476].

L'annotation GENCODEv35lift37 pour hg19 et l'annotation GENCODEv35 pour hg38 sont largement similaires, avec 95 % des gènes présents dans les deux annotations et à l'exception de 3 515 et 1 764 gènes qui n'ont été annotés que dans hg19 et hg38 respectivement. Après exclusion du chromosome Y, 58 373 gènes ont été annotés dans les deux annotations, dont 96,8 % avec des modèles identiques. Sur les 1 844 gènes présentant des différences de modèle, la plupart avaient des différences dans le nombre d'exons annotés (723 gènes) ou des différences dans la longueur d'un ou plusieurs exons constitutifs (980 gènes). Parmi les gènes avec des modèles identiques, la majorité avait également des séquences exoniques identiques, et seulement 1 199 différences ont été expliquées par de petites variations dans la séquence exonique. Notons que 40 gènes avaient une similarité de séquence anormalement faible (similarité de Jaro-Winkler ≤ 0,5), et qu'ils se trouvaient tous sur le chromosome X.

De même, environ 94 % (59 815/63 710) des gènes étaient présents à la fois dans l'annotation GENCODEv35 pour hg38 et dans l'annotation GENCODEv35 CAT/Liftoff v2 pour CHM13. 322 gènes étaient spécifiques de hg38 et 3 573 gènes présents dans l'annotation CHM13 n'avaient pas de modèle de gène correspondant dans hg38, y compris 2210 nouveaux gènes dans des régions non synténiques. Approximativement 24 % (13 929) des gènes présents dans les deux annotations présentaient des différences dans le modèle génique. En outre, 30 % (17 660) des gènes dont les modèles étaient identiques présentaient des différences dans la séquence génétique sous-jacente, ce qui peut être en partie imputable à l'utilisation de lignées cellulaires différentes pour construire les assemblages hg38 et CHM13. Seuls deux gènes autosomiques présentaient cependant une réelle faible similarité de séquence (CDC27P1 et ENSG00000279501) entre hg38 et CHM13.

Ces résultats, disponibles sous la forme d'une table d'équivalence, soulignent la cohérence de l'annotation de GENCODEv35 pour hg38 et de ses homologues pour hg19 et CHM13. Ces annotations ont ainsi été considérées comme suffisamment comparables, laissant possible

l'étude des effets de l'assemblage indépendamment de l'annotation pour la majorité des gènes.

Impact des gènes exprimés, dont les modèles sont spécifiques d'une unique annotation, sur les interprétations biologiques

En outre, 169 gènes spécifiques de l'annotation hg19 étaient exprimées et 14 pour l'annotation hg38 dans la comparaison hg19:hg38. Aucun n'était associé à une maladie connue. La majorité (92 %) de ceux spécifiques de l'annotation hg19 se trouvaient dans des régions connues comme problématiques et seulement 33 % étaient des PCG ou lncRNA. En revanche, seuls 14 % des gènes exprimées spécifiques de hg38 chevauchaient une région problématique et 88 % étaient des PCG ou lncRNA. Cela suggère que hg19 présente davantage de bruits parasites que hg38 et que la majorité des gènes perdus lors du passage de hg19 à hg38 sont probablement d'une pertinence clinique limitée.

Dans la comparaison hg38:CHM13, nous avons détecté 68 gènes exprimées spécifiques de l'annotation hg38 et 335 pour CHM13, dont 60 % et 44 % étaient des PCG ou lncRNA, respectivement. La majorité (67 % pour hg38 et 78 % pour CHM13) se trouvait dans des régions connues pour être problématiques et ces gènes doivent donc être considérés avec prudence.

Pour illustrer, les gènes CFHR1 et CFHR3 liés au syndrome hémolytique et urémique atypique [477, 478] sont modélisés pour hg19 et hg38 et exprimés, mais absents pour CHM13. Ce résultat serait lié à l'usage d'une seule lignée cellulaire pour l'assemblage CHM13, contrairement à la diversité employée pour hg38. De plus, l'étude de Hamza et al. [479] rapporte que la détection des variants structurels à l'origine de la maladie n'a pas été possible lors de l'alignement sur CHM13, même avec un séquençage *long-read*, ce qui suggère que ces gènes ne devraient pas être évalués à l'aide de CHM13.

D'autre part, un gène spécifique de hg38, SIK1, lié au développement de l'encéphalopathie épileptique et développementale [480], présente une duplication SIK1B dans cette même annotation. Cette duplication n'est pas présente chez CHM13. Par conséquent, SIK1B a une expression 5,5 fois plus élevée dans CHM13 par rapport à hg38 en s'accaparant les *reads* de SIK1, les deux ayant des séquences proches du fait de leur paralogie.

Identification des gènes présents dans plusieurs assemblages et avec une expression significativement différente entre annotation

Les gènes annotés dans plusieurs annotations peuvent également présenter des différences d'expression malgré l'usage du même échantillon biologiques. Ces gènes sont alors qualifiés de « différentiellement quantifiés » et notés DQ. Pour identifier ces gènes, seuls ceux ayant avec une expression > à 0,1 TPM dans au moins 30 % des échantillons testés dans les deux versions ont été conservés.

L'ensemble des tissus a permis de tester la quantification différentielle pour 31 275 gènes entre hg19:hg38 incluant 78% des PCG connus. Au total, 202 gènes (donc 94 PCG) présentaient des différences significatives et importantes (*abs(logFC)* > 1) pour leur expression entre hg19 et hg38 dans au moins un tissu. Bien que le nombre de gènes DQ variait selon les tissus, 125 l'étaient dans plus d'un tissu. Notamment, 23 gènes exprimés dans les six tissus ont montré une quantification différentielle dans tous les tissus, ce qui est cohérent avec l'hypothèse qu'un changement d'annotation soit détecté dans tous les tissus dans lesquels un gène est suffisamment exprimé. La majorité des gènes DQ (180/202, 90 %) chevauchaient en réalité des régions erronées ou difficiles à séquencer dans hg19 ou hg38 [471, 481]. Les modifications de la séquence génique sous-jacente et/ou du modèle génique expliquent en partie les différences pour 18 des 22 gènes qui ne chevauchaient pas des zones du génome avec des problèmes connus.

Des résultats semblables ont été observés entre hg38 et CHM13 où 1 341 gènes étaient considérés comme DQ dans au moins un tissu. La majorité des gènes DQ (1028/1341) étaient en réalité observée dans plus d'un tissu et 452 l'étaient dans les six tissus disponibles. Cependant, seul 38% des gènes DQ étaient dans régions erronées ou compliquées à séquencer et 52% des restants présentaient des modifications dans leurs modèles géniques ou étaient localisés dans les zones de changement d'assemblage.

Pour finir, notons que dans les deux comparaisons, 267 gènes DQ sont impliqués dans une maladie rare connue dans la base de données OMIM [482] (7 pour hg19:hg38 et 262 pour hg38:chm13).

A titre d'exemple, le gène présentant le f*old-change* d'expression maximum entre hg19 et hg38 est SCN8A, qui a été associé à l'encéphalopathie épileptique à début précoce [483], avec une expression 83x plus élevée dans les iPSC dans hg38 par rapport à hg19. Pour hg38 et

CHM13, le gène est SH2B3, dont l'expression est 1097 fois plus élevée dans hg38, et qui est associé à l'érythrocytose somatique.

Identification des gènes annotés dans plusieurs annotations, mais exprimés dans une seule d'entre elles

La quantification différentielle ne peut évaluer l'impact du choix de la construction que pour les gènes qui sont annotés et suffisamment exprimés dans les deux constructions. Ainsi, nous avons également étudié les gènes exclus de l'analyse de quantification différentielle en raison de niveaux d'expression insuffisants dans une seule construction malgré leur présence dans les deux annotations. En comparant hg19:hg38, 96 gènes quantifiés uniquement dans hg19 et 126 uniquement dans hg38 ont été détectés et plus des 2/3 de ces 222 gènes chevauchaient des régions erronées ou problématiques. Dans la comparaison hg38:CHM13, 309 gènes mutuellement annotés ont été détectés exclusivement à partir de l'alignement hg38 et 387 à partir de l'alignement CHM13, dont 39 % et 54 % se trouvaient dans des régions erronées ou problématiques, respectivement. Ainsi, une grande prudence est de mise lors de l'étude des gènes dont l'expression est exclusive à un seul assemblage.

Prenons l'exemple de BMS1P8 dont l'augmentation d'expression a été liée à des taux de survie plus faibles pour le carcinome hépatocellulaire sur la base des niveaux d'expression observées chez hg19 [484]. En réalité, cette association semble être un artefact dû à une erreur dans cette région dans hg19. Bien que BMS1P8 soit annoté à la fois dans hg38 et CHM13, les *contigs* utilisés pour construire la région ont été mis à jour dans l'assemblage hg38 [481] et l'expression de ce gène n'est plus détectée que dans hg19.

Les changements apportés au modèle d'un gène entre les versions peuvent également avoir un impact lors de l'alignement sur le transcriptome conduisant alors à une expression exclusive de la version. Illustrons avec le gène codant pour la protéine PDGFRB qui est impliqué dans de nombreuses maladies rares, dont le syndrome de surcroissance de Kosaki [485, 486]. Dans le sang, 194 échantillons sont considérés comme l'exprimant pour hg38, mais aucun pour CHM13. En réalité, l'annotation CHM13 pour PDGFRB est plus complexe que les modèles de gènes hg38 et hg19, avec trois transcrits supplémentaires, ce qui entraîne des taux de multimappage plus élevés et rend la quantification précise plus difficile lors de l'alignement basé sur le transcriptome. Par conséquent, lors de la mise en

correspondance avec des modèles de gènes complexes, une méthode de quantification utilisant un alignement basé sur le génome pourrait s'avérer plus appropriée.

1.3.3. Valorisations associées

Ces travaux ont fait l'objet :

- d'un article en finalisation d'écriture : Ungar R*, Goddard P*, Jensen T, **Degalez F**, Smith K, Jin C, Bonner D, Bernstein J, Wheeler M, Montgomery S. (2023). Impact of genome build on RNA-seq interpretation and rare disease diagnosis. – En finalisation d'écriture. **Cet article sera soumis à Nature Genetics. Pour des raisons de confidentialité, la dernière version du papier n'est pas reproduite ci-après.**

- d'une communication orale : Goddard P, Ungar R, Jensen T, Marwaha S, Bonner D, **Degalez F**, Smith K, Montgomery S. (2022). Genome reference impacts RNA-seq interpretation and rare disease diagnosis. Exposé oral lors de l'"American Society of Human Genetics" (ASHG), Los Angeles, California, United-States

d'un poster : Goddard P, Ungar R, Jensen T, Marwaha S, Bonner D, **Degalez F**, Smith K, Montgomery S. (2022). Genome reference impacts RNA-seq interpretation and rare disease diagnosis. Poster présenté lors de l'"American Society of Human Genetics" (ASHG), Los Angeles, California, United-States

1.4. Mise en place d'un *pipeline* de détection de lncRNA orthologues entre différentes espèces *(Résumé de travaux)*

1.4.1. Contexte et objectifs

Comme présenté dans l'introduction générale §1.3, l'identification des relations d'orthologie permet de transférer des connaissances fonctionnelles entre espèces par extrapolation et permet soit de s'affranchir, soit d'orienter une partie des démarches expérimentales afin de compléter l'annotation des génomes. Cependant, à ce jour, les études comparatives se sont particulièrement intéressées aux gènes codant des protéines (PCG) en analysant la conservation des séquences [243, 245], la synténie [246, 247] ou encore les arbres phylogénétiques [250, 251]. Des bases de données généralistes comme Ensembl [258, 259] recensent par ailleurs ces PCG homologues.

Cependant, les techniques de séquençage haut-débit ont permis de mettre en évidence de nouvelles classes de gènes tels que les ARN longs non-codants (lncRNA) aux fonctions régulatrices diverses (voir Introduction §1.2.3). L'étude des lncRNA par comparaison génomique apparait difficile, car ces gènes sont très peu conservés en termes de séquence primaire [276, 283]. Seuls quelques unités de la séquence primaire seraient conservées afin de maintenir les aspects fonctionnels. Plus précisément, la conservation de la séquence semble se faire par blocs de 5 à 30 nucléotides, les k-mers [487]. Face à cette problématique, seules quelques bases de données ont actuellement identifié des homologies entre lncRNA comme SyntDB [488] qui a identifié des homologies pour le groupe des primates et NONCODE v6 pour une douzaine d'espèces de différentes branches, incluant la poule [112].

Dans ce contexte, nous avons mis en place un *pipeline* combinant trois méthodes qui peuvent être utilisées pour toutes les espèces d'intérêt : la méthode n°1 utilise un triplet "PCG-lncRNA-PCG" avec 2 PCG orthologues de part et d'autre du lncRNA comme "ancre". La méthode n°2 est basée sur des paires "lncRNA-PCG" avec un seul PCG orthologue et sur la comparaison des configurations respectives entre les lncRNA et PCG dans les deux espèces. Enfin, la dernière méthode considère l'alignement des lncRNA en utilisant la méthode d'alignement multiple des génomes "Mercator-Pecan". Appliquées sur 11 espèces couvrant une large échelle phylogénétique, des mammifères à la poule, ces trois méthodes apparaissent complémentaires quelle que soit la paire d'espèces considérée. En s'appuyant

sur les 18 805 lncRNAs humains (Ensembl V106), environ 9 000 lncRNA sont identifiés avec un orthologue chez la poule, dont environ 3 000 détectés par deux méthodes et 1 000 détectés par les trois méthodes respectivement.

1.4.2. Matériels et démarches

Choix des espèces d'études – Dans le cadre de l'analyse et du développement du *pipeline*, 11 espèces ont été étudiées. L'accent a été mis sur les espèces domestiques à savoir : *i)* le chien - *Canis lupus famliaris – CanFam3.1* ; *ii)* la vache – *Bos taurus* – ARS-UCD1.2 ; *iii)* le cheval – *Equus caballus – EquCab3.0* ; *iv)* le porc – *Sus scrofa – Sscrofa11.1* ; *v)* la chèvre – *Capra hircus – ARS1* et *vi)* la poule – *Gallus – GRCg6a*. Afin d'enrichir le groupe des oiseaux, *vii)* la dinde – *Meleagris gallopavo – Turkey_5.1*, ainsi que le *viii)* diamant mandarin – *Taeniopygia guttata – bTaeGut1_v1.p* ont été intégrés dans les analyses. De même, en raison de leur statut de référence et de leur annotation plus complète, *ix)* le poisson-zèbre – *Danio rerio – GRCz11* ; *x)* la souris – *Mus musculus* – GRCm39 et *xi)* l'humain – *Homo sapiens – GRCh38.p13* ont également été inclus. Les fichiers d'annotations (GTF) utilisés sont ceux produits par la base de données Ensembl (v.104). D'autre part, l(annotation enrichie en lncRNA produit par Jehl et al. [124] a été utilisé pour la poule.

Nomenclature pour l'orthologie des lncRNA – Basé sur la classification classique employée dans les études d'orthologies entre gènes, six sous-cas peuvent ici être considérés pour les lncRNA :

- *"one_to_zero"* : un lncRNA du génome source n'a pas de lncRNA équivalent identifié dans le génome cible.
- "*many_to_zero*" : plusieurs lncRNA du génome source n'ont pas de lncRNA équivalent identifié dans le génome cible.
- "*one_to_one*" : un lncRNA du génome source a un unique lncRNA équivalent identifié dans le génome cible.
- "*many_to_many*" : plusieurs lncRNA du génome source ont plusieurs lncRNA équivalents identifiés dans le génome cible.
- "*many_to_one*" : plusieurs lncRNA du génome source ont un unique lncRNA équivalent identifié dans le génome cible.
- "*one_to_many*" : un lncRNA du génome source a plusieurs lncRNA équivalents identifiés dans le génome cible.

Les méthodes présentées considèrent deux espèces, l'une comme source, l'autre comme cible. La figure « *Method* » du poster associé illustre les trois méthodes détaillées ci-après :

Méthode 1 – Conservation des triplets « PCG-lncRNA-PCG » – Pour l'espèce source et cible, chaque lncRNA est localisé et les deux PCG voisins les plus proches sont identifiés. La position relative du lncRNA par rapport aux deux PCG est conservée pour des analyses additionnelles. À l'aide de la base de données BioMart [259], et pour chaque triplet « PCG-lncRNA-PCG », les PCG voisins du lncRNA dans l'espèce source sont mis en parallèle de ceux de l'espèce cible pour vérifier leur lien d'orthologie. Seuls les cas où les deux PCG ont un orthologue "one_to_one" sont analysés. Par la suite, le(s) lncRNA contenu(s) entre les deux PCG du génome source est/sont classifié(s) par rapport à ceux présents dans le génome cible suivant la classification énoncée précédemment. Pour finir, l'orientation respective des gènes entre les deux génomes est analysée : elle peut être, *i)* identique, si l'orientation des gènes est la même pour les deux espèces, *ii)* inverse, si les gènes sont tous en orientation opposée par rapport au génome de référence ou encore *iii)* discordant, si les orientations entre génome source et cible ne suivent pas de lien logique.
Pour les lncRNA dans le cas "*many*", une étape consistant à regrouper les gènes consécutifs avec une orientation identique peut être appliquée.

Méthode 2 – Conservation des configurations « lncRNA-PCG » – Pour l'espèce source et cible, la configuration et la distance de chaque lncRNA par rapport au PCG le plus proche est calculé en utilisant FEELnc [122]. À l'aide de la base de données BioMart [259] et pour chaque couple « lncRNA-PCG », le PCG dans l'espèce source est mis en parallèle de celui de l'espèce cible pour vérifier leur lien d'orthologie. Seuls les cas où le PCG a un orthologue "*one_to_one*" sont conservés. Ensuite, les couples « lncRNA-PCG » de l'espèce source sont confrontés aux couples de l'espèce cible pour observer si les configurations sont conservées. Afin de prendre en compte la variabilité de la qualité des annotations, la classification des configurations peut être assouplie selon cinq niveaux :

- "*strict*" : configurations telles que fournies par FEELNc (*e.g.*, lincSSdw_n.1.n)
- "*inter1*" : suppression des informations complémentaires sur l'association des transcrits (*e.g.*, lincSSdw_n.1.n→ lincSSdw)

- "*inter2*" : les lncRNA sont étiquetés comme géniques (lncg) ou intergéniques (linc) et comme convergents, divergents/antisens ou même brin (*e.g.*, lincSSdw_n.1.n → lincSS)
- "*open1*" : les lncRNA sont étiquetés comme convergents, antisens ou brins identiques (*e.g.*, lincSSdw_n.1.n → SS).
- "*open2*" : les lncRNA sont soit étiquetés comme antisens, soit comme même brin. (*e.g.*, lincConv → AS)

Par la suite, le(s) lncRNA contenu(s) dans chaque couple « lncRNA-PCG » du génome source est/sont classifiés par rapport à ceux présent dans le génome cible suivant la classification énoncé précédemment. Pour les lncRNA dans le cas "*many*", une étape consistant à regrouper les gènes consécutifs avec une orientation identique peut être appliquée.

Méthode 3 – Méthode d'alignement – Le groupe "*63 amniota vertebrates*" de la base de données Compara [258] a été utilisé pour extraire les informations issues de la méthode d'alignement multiple des génomes "*Mercator-Pecan*" [489].Ce groupe contient toutes les espèces utilisées dans cette étude, à l'exception du poisson-zèbre. Rapidement, ces méthodes construisent d'abord des cartes synténiques entre les génomes puis des alignements dans ces régions, ce qui facilite la détection des lncRNA même si la séquence n'est pas bien conservée. Trois cas peuvent être envisagés :

- Le lncRNA de l'espèce source s'aligne dans l'espèce cible sur un modèle existant
- Le lncRNA de l'espèce source s'aligne dans l'espèce cible mais aucun modèle de gène n'existe
- Le lncRNA de l'espèce source ne s'aligne pas dans l'espèce cible

Les trois méthodes présentées peuvent être appliquées indépendamment, mais leur intérêt réside dans leur utilisation conjointe.

Réseaux de co-expression – Pour un couple de lncRNA potentiellement orthologue entre la poule et l'humain, la co-expression du lncRNA avec les autres PCG a été observée pour chacune des espèces. Les 50 PCG les plus co-exprimées ont alors été sélectionnés et le nombre de PCG orthologues entre les deux ensembles a été quantifié. Les données d'expressions

proviennent du projet GTEx humain [389] incluant 52 tissus et de l'atlas GRCg6b de la poule enrichie en lncRNA [124].

1.4.3. Résultats préliminaires

Les analyses d'orthologie ont été menées pour 11 espèces telles qu'évoquées dans le matériel et méthode. Cependant, l'humain a ici été considéré comme espèce source de par la qualité de son annotation.

Comme le montre la Table 1 (voir Poster), le nombre de gènes identifiés dans chaque espèce est variable. Cette variabilité apparait très importante pour les lncRNA avec un minimum à 1 034 pour la dinde, un maximum à 17 734 pour l'humain et un total de 24 835 lncRNA pour l'annotation enrichie du génome de la poule. Si cette variabilité est moins importante pour les PCG, elle peut tout de même être observée avec un minimum de 16 226 modèles pour la dinde et un maximum de 25 432 pour le zebrafish. Notons que cette variabilité à son importance, car les PCG et leurs liens d'orthologies sont utilisés comme « ancres » dans les deux premières méthodes. Pour chacune des 10 espèces, quelques dizaines à quelques centaines de lncRNA semblent orthologues avec l'humain selon la méthode n°1.

Concernant la méthode n°2, le nombre de lncRNA potentiellement orthologues avoisinent les quelques milliers, soit plus de 25 % des lncRNA identifiés dans l'espèce cible. Notons tout de même des chiffres plus élevés pour la souris, deuxième espèce la mieux annotée après l'humain. Pour finir, la méthode n°3 met en évidence plusieurs milliers d'alignement pour les lncRNA humain chez les espèces cibles. Pour les espèces phylogénétiquement proches de l'humain, incluant notamment les mammifères, en moyenne 8 000 lncRNA humain présentent un résultat d'alignement, ce chiffre avoisinent les 4 800 pour les espèces plus éloignées comme la poule et le diamant mandarin. Certaines ne présentent pas de résultat, car elles n'étaient pas prises en compte par Compara pour l'approche d'alignement par Mercator-Pecan. Si chacune de ces méthodes permettent de caractériser individuellement les lncRNA humain et leurs liens d'orthologies potentielles, leur considération simultanée permet de parfaire les analyses. Ainsi, l'union des méthodes a permis d'identifier plus de 9 000 lncRNA orthologues entre l'humain et l'espèce d'intérêt, même si le modèle n'était pas annoté (intérêt de la méthode n°3). De manière intéressante, alors que ce nombre avoisine les 6 000

pour le diamant mandarin, plus de 9 000 lncRNA potentiellement orthologues, un nombre équivalent aux espèces phylogénétiquement proches, sont identifiés chez la poule, ce qui souligne la force de l'annotation enrichie.
De manière intéressante, 23 lncRNA ont été identifiés chez l'humain comme potentiellement orthologues chez neuf autres espèces avec, par exemple, INST6-AS1 (voir Figure 1, Poster), lncRNA en antisens de INST6 et partageant potentiellement un promoteur et une fonction commune.

Cependant, afin d'appuyer davantage les liens d'orthologie présumés, l'intersection des méthodes apparait plus intéressante. Ainsi comme illustré dans la Figure 2 (voir Poster), et comme attendu, les méthodes n°1 et n°2, de par leur nature, tendent à se recouvrir davantage et ce peu importe le couple d'espèce considéré, ici humain:poule (HSA:GGA) et souris:poule (MMU:GGA). À la suite de ces observations, sur les 9 142 lncRNA orthologues HSA:GGA détectés par au moins une méthode, 3 609 (39 %) ont été sélectionnés et considérés comme plus fiables, car trouvés par au moins deux méthodes (2 765 cas) ou par la méthode n°1 ou n°2 sécurisées par deux méthodes pour GGA:MMU (1 873 cas). Plusieurs de ces lncRNA orthologues étaient spécifiques de tissus et avaient des gènes co-exprimés associés aux mêmes termes enrichis, liés à la fonction du tissu, comme illustré en Figure 4 (voir Poster).

1.4.4. Limites

Si la complémentarité des approches présentées permet d'apporter des arguments sur de potentiels liens d'orthologie entre les lncRNA de différentes espèces, à ce stade du projet, plusieurs limites font déjà surface. Tout d'abord, notons que toutes les analyses sont dépendantes des annotations des gènes qui peuvent être de qualité très variable selon les espèces considérées. En effet, le nombre de lncRNA identifiés apparait sous-évalués et la fiabilité des modèles est encore discutable pour la majorité des espèces. Ainsi, si aucun lien d'orthologie n'est identifié, l'absence d'annotation du gène en question dans l'espèce étudié peut-être la raison principale. Cette conclusion peut en partie être nuancée par l'application de la méthode n°3 d'alignement. Cependant, a contrario, la seule présence d'une équivalence par cette méthode n'induit pas la présence systématique d'un modèle génique fonctionnel.

Dans tous les cas, les méthodes employées doivent ne servir que de bases pour générer des hypothèses qui devront être testées expérimentalement. Toujours concernant l'annotation, il est maintenant connu que les gènes évoluent d'un point de vue fonctionnel et, de manière simpliste, PCG, lncRNA et pseudogènes peuvent provenir d'un même gène ancestral commun. Cependant, dans notre cas, l'identification des gènes se fait selon leur biotype, les méthodes mis en place ne prennent donc pas en compte l'évolution des gènes et l'identification par exemple d'un lncRNA orthologue d'un PCG n'est pas envisagé.

Une autre limite concerne le maintien à jour des résultats obtenus. En effet, les annotations et les assemblages de l'ensemble des espèces évoluent, ce qui peut changer de manière considérable les résultats obtenus. Pour exemple, le *pipeline* a été appliqué à la fois sur l'annotation usuelle du génome de la poule fournie par Ensembl (GRCG6a, V104) mais également en utilisant l'annotation enrichie telle que présentée dans Jehl et al. [124] et les résultats observés apparaissaient alors différents. D'un point de vue computationnelle, cela nécessiterait de réitérer les analyses pour chaque nouvelle version d'annotation de n'importe quelle espèce, ce qui représente un temps de calcul trop important.

Pour finir, la principale faiblesse de ce *pipeline* repose sur sa dépendance aux bases de données externes et notamment Ensembl et ses outils dérivés. Ainsi, particulièrement pour les PCG, il n'est possible d'utiliser que des identifiants respectant la nomenclature Ensembl (ENSxx). Même si des équivalences existent entre les identifiants géniques de différentes bases de données, comme RefSeq et Ensembl par exemple, ils ne correspondent pas de manière exacte aux mêmes modèles géniques, ce qui peut biaiser les analyses. De manière plus contraignante, certains modèles géniques peuvent exister dans une seule base de données et les relations d'équivalences sont alors inexistantes. De même, dans la méthode n°3, seul un ensemble d'espèce spécifique peut être étudié. En effet, les résultats issus de Compara pour l'alignement Mercator-Pecan ne correspondent qu'à un ensemble d'espèces prédéfinies, rendant impossible les analyses pour les espèces en dehors ce cadre.

1.4.5. Perspectives

Améliorations techniques – Si ces travaux présentent déjà quelques résultats convaincants, les limites évoquées dans le paragraphe précédent doivent être pris en compte. Il est ainsi envisagé d'automatiser la conversion des identifiants géniques, quelle que soit la source, en identifiant Ensembl. Si pour les identifiants RefSeq, des tables d'équivalences peuvent être extraites de BioMart, pour les autres, il est envisagé d'employer une méthode proche de celle évoquée dans l'article §1.1, c'est-à-dire en considérant deux modèles géniques comme équivalents s'ils se chevauchent à un seuil fixé. En effet, il n'est pas possible de s'affranchir des listes de PCG orthologues fournis par BioMart, cette conversion apparait alors comme une solution. Concernant la méthode n°3 par alignement Mercator-Pecan, il faudrait pouvoir appliquer la même démarche que celle mise en place par Ensembl mais pour un ensemble d'espèces incluant toutes nos espèces à l'étude.

Afin d'apporter des indices supplémentaires venant appuyer davantage les liens d'orthologie, la mise en place de méthodes supplémentaires est envisagée. Ainsi, au niveau de l'alignement, il serait possible d'intégrer des logiciels spécialisés tels que SEEKR [487] ou LncLOOM [286]qui reposent sur l'identification de combinaisons de motifs courts dans des séquences supposées homologues de différentes espèces. Pour finir, et même si une ébauche de cette approche a déjà été évoquée en résultats, il serait intéressant d'intégrer les profils d'expression géniques, que ce soit pour les comparer entre gènes supposés orthologues, mais également pour voir si ces gènes semblent appartenir à des mêmes réseaux de co-expression tels que fournis par WGCNA. Cependant, bien que cette approche se rapproche du fonctionnel, le peu de données d'expression disponible pour certaines espèces se présente comme un frein majeur.

Améliorations des informations biologiques – En se focalisant notamment chez la poule, la sortie du nouvel assemblage GRCg7b et la parution du nouvel atlas enrichi en lncRNA (voir Résultats §1.1) [125] comportant une quantité plus importante de modèles que la version précédente sous GRCg6a pousse à relancer les analyses avec les dernières données disponibles. Les autres génomes ayant également connu des améliorations (*e.g*, la sortie du T2T humain ou encore la parution du nouveau génome bovin), il apparait opportun d'intégrer ces avancées. Les lncRNA identifiés comme conservés chez la poule feront spécifiquement l'objet d'analyses supplémentaires d'un point de vue expérimental. En effet, un sous-ensemble, sélectionné selon des critères précis tels que leur position par rapport à des PCG

d’intérêt et/ou leur seuil d’expression, sont prévus pour faire l’objet de capture afin d’être plus finement annotés par la suite.

L’objectif est donc de poursuivre ces travaux en proposant un *pipeline* d’analyse complet permettant de faciliter les analyses par orthologie pour les lncRNA et d’opérer des vérifications biologiques. Ces travaux pourraient conduire à la publication d’un article scientifique.

1.4.6. Valorisations associées

Ces travaux ont fait l'objet :

- d'un poster présenté à deux congrès :

    - **Degalez F**, Lagoutte L, Lecerf F, Vlach M, Lagarrigue S. (2022). Gene orthology detection for long non-coding RNA. Poster présenté aux "Open Day of Computational Biology and Mathematics" (JOBIM), Rennes, France.

    - **Degalez F**, Allain C, Lagoutte L, Lagarrigue S. (2023). Gene orthology detection for long noncoding RNA (lncRNA). Poster présenté dans la session spécialisée "Animal Epigenetics" au 39ème congrès de l'"International Society for Animal Genetics" (ISAG), Cape Town, South Africa. **Ce poster est reproduit ci-après ;**

GENE ORTHOLOGY DETECTION FOR LONG NON CODING RNA (lncRNA)

F. Degalez, L. Lagoutte, F. Lecerf, C. Allain, S. Lagarrigue

fabien.degalez@inrae.fr PEGASE, INRAE, Institut Agro, 35590 Saint Gilles, France

CONTEXT

→ Long non-coding RNAs (lncRNAs), with at least 200 nt & low potential coding capabilities are a major component of regulatory elements in genomes. More than 15,000 loci have been modeled in human.

→ However, the role of most of them remains to be clarified.

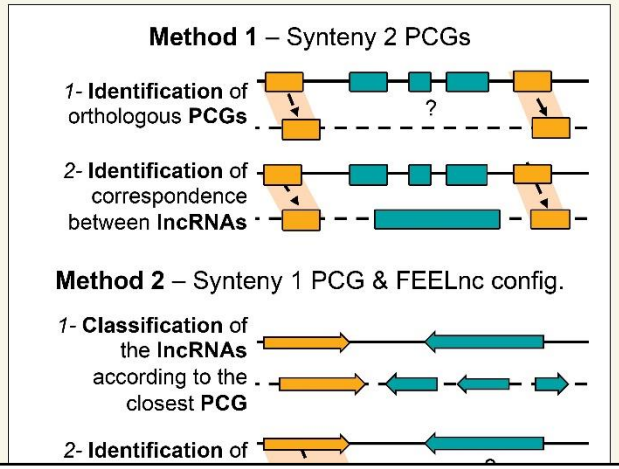
• Exploring lncRNA conservation between species is an approach to strengthen the annotation of lncRNAs by inferring function in one species from another one more studied such as human or mouse, as has been done previously for protein coding genes (PCGs).

• However, unlike PCGs, lncRNA sequences are not well conserved across species [1]. Therefore, no lncRNA orthologs are reported in reference databases such as Ensembl BioMart, regardless of species.

• In this context, we have developed a workflow combining 3 approaches (Method 1, 2 and 3) that can be used for any species of interest and have applied it on 11 species covering a large phylogenetic scale from mammals to chicken.

METHOD

Gitlab – *f.degalez/lncrna_orhtologfinder*

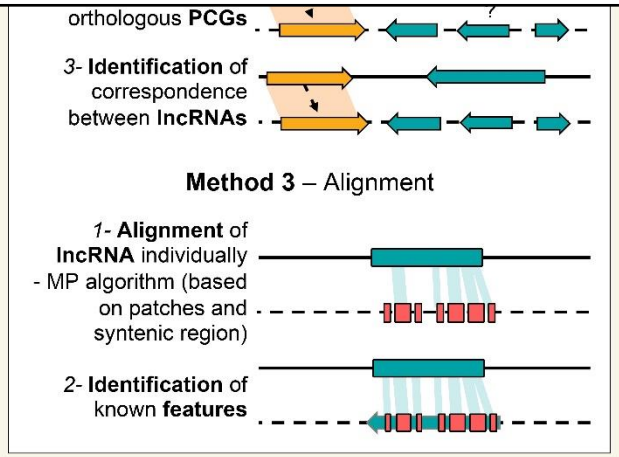


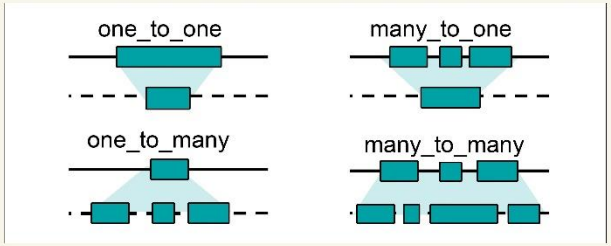
RESULTS - NUMBERS

1 ***Table 1*: Orthologous lncRNAs between HSA & 10 species (mammals, birds & fish).**

** The reference Ensembl chicken gene atlas is enriched in lncRNA loci [2]*

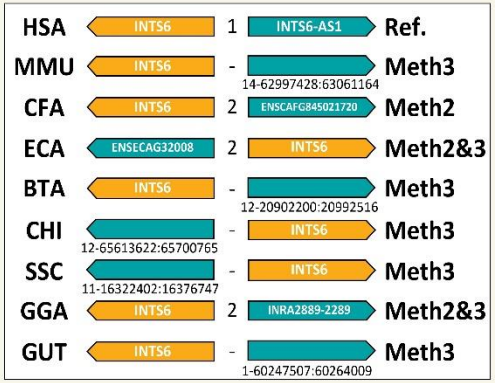
Source : Human Total : 61552 - PCG : 19994 - lncRNA : 17734												
				method 1			method 2			method 3	
Total	PCG	lncRNA	Target	*1to1*	*many*	Tot	*1to1*	*many*	Tot	Match	U
55414	21884	9949	***Mouse***	457	876	1333	1214	3190	4404	8660	**10867**
31121	20500	6479	***Dog***	112	230	342	665	1543	2208	-	**2372**
30371	20912	7241	***Horse***	144	515	659	746	1775	2521	7684	**9216**
27607	21861	1480	***Cow***	65	99	164	260	480	740	8947	**9342**
27271	21343	2705	***Goat***	73	171	244	319	821	1140	8824	**9378**
31908	21280	6790	***Pig***	150	343	493	735	1785	2520	8885	**10205**
45289	17859	24835	**** Chicken***	244	858	1102	1368	4955	6323	4816	**9142**
17970	16226	1034	***Turkey***	18	37	55	137	275	412	-	**437**
22150	16619	4757	***Zebrafinch***	60	213	273	548	1221	1769	4835	**6099**
32520	25432	2222	***Zebrafish***	14	23	37	156	355	511	-	**529**





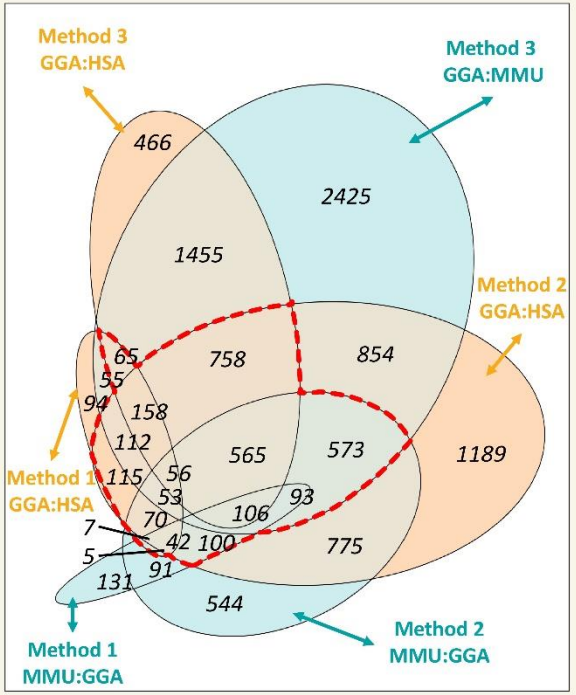
2 *Fig. 1*: One of the 23 lncRNAs in HSA identified as orthologous in 9 species by at least one method.

Orthologous transcripts are either antisense (1) or divergent (2) with respect to the closest PCG, suggesting partial knowledge of TSS position in some species. LncRNA capture combined with long-read sequencing is in progress.



3 *Fig. 2*: Subset of 3609 orthologous HSA:GGA lncRNAs considered more reliable by combining several methods.

*Out of the **9142** HSA:GGA orthologous lncRNAs detected by at least 1 method (Table 1), **3609 (39%)** lncRNAs were selected (red section) as more reliable because found orthologous by at least 2 methods (2765 cases) or by the method 1 or method 2 secured by 2 methods for GGA:MMU (1873 cases)*



RESULTS – FUNCTIONAL ANNOTATION

4

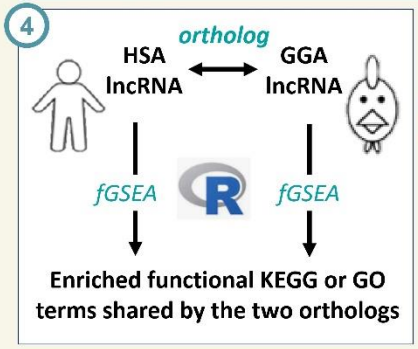
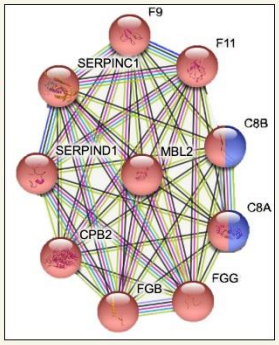


Fig. 4: Several tissue-specific orthologous lncRNAs had co-expressed genes associated to the same enriched terms related to the tissue function.

Here, the case of liver-specific lncRNA orthologs in GGA and HSA associated to "Complement and coagulation cascades" (10 / 82); FDR: 1.29 e-14.



The results presented here consider annotations and/or genome assemblies that may have been updated, such as the enriched annotation for the chicken genome (see poster P63). New analyses are currently in progress to include the major modifications.

Acknowledgments: We thank Thomas Derrien (CNRS, IGDR, Rennes), Sylvain Foissac (INRAE, UMR GenPhySe, Toulouse) and Hervé Acloque (INRAE, UMR GABI, Paris) for discussions.

References: 1 – T. Noviello et al. BMC Bioinformatics (2018) 19, 407. doi: 10.1186/s12859-018-2441-6
2- F Jehl et al. Sci Rep. (2020) 10(1):20457. doi: 10.1038/s41598-020-77586-x

INRAE L'INSTITUT agro ISAG

2. Identification de SNP avec des génotypes fiables

2.1. Utilisation des données RNAseq pour la détection de SNP avec des génotypes fiables et exemples d'applications *(Résumé d'article)*

2.1.1. Contexte et objectifs

Le séquençage ARN (RNAseq) est devenu en une dizaine d'années la technologie de référence pour l'étude transcriptomique, en remplacement des puces à ADN [290]. Cette approche permet d'étudier les profils d'expression génique dans une grande variété d'organismes, afin de mieux comprendre les mécanismes génétiques sous-jacents à divers phénotypes, maladies ou réponses environnementales [490–492] (voir Introduction §2.1). En effet, depuis ces dernières années, une quantité importante de données RNAseq est disponible dans les bases de données publiques du fait d'un nombre croissant d'études portant sur différentes populations et conditions [493]. Cependant, de par son étape de séquençage, le RNAseq peut permettre à lui seul de détecter des variations génomiques dans les régions exprimées, comme l'a montré Piskol et al. en 2013 [337]. Cette approche présente alors de nombreux avantages, notamment dans les espèces non-modèles pour lesquelles peu de données de séquençage ADN (DNAseq) existe. Tout d'abord, le RNAseq cible les régions codantes pouvant avoir un impact fonctionnel sur la fonction protéique [494, 495]. De plus, les régions transcrites sont nombreuses et bien réparties dans le génome comme souligné par le consortium ENCODE [99]. Finalement, grâce au RNAseq, si un gène possède un SNP hétérozygote, il est possible d'identifier des différences d'expression significatives entre le chromosome paternel et maternel, on parle alors d'expression allèle-spécifique (ASE pour *Allele Specific Expression*) [352].

Malgré ces avantages, la détection de SNP par RNAseq présente trois défis techniques majeurs : *i)* l'alignement des *reads* est complexe du fait de l'épissage des transcrits matures [496] ; *ii)* le nombre de *reads* à une position est très variable selon l'expression des gènes entrainant des variabilités entre SNP et individus dans les profondeurs et les génotypages par rapport au DNAseq 20X (voir Figure 1) [497] ; *iii)* des mécanismes d'édition d'ARN changeant la séquence existante au niveau de l'ADN génomique peuvent introduire des variations absentes au niveau de l'ADN [341]. Cependant, ce dernier cas est assez rare, peu d'événements d'édition sont généralement identifiés que ce soit chez la souris [344], la poule [345–347] (<200 événements) ou l'homme [348, 349] (<1000 événements). L'édition

d'ARN est plus fréquente dans les régions répétées et donc peu représentée dans les *reads* alignés [350]. Au final, peu d'études ont utilisé le RNAseq pour la détection de variants et le génotypage depuis 2013 [498–504].

L'objectif de ce travail est de mettre en place une procédure de détection de SNP et de génotypage, à partir de données RNAseq, en l'appliquant pour la poule. Dans un premier temps, la fiabilité des SNP détectés par RNAseq est évaluée par comparaison à ceux obtenus par DNAseq 20X, considérés comme la vérité, sur les mêmes échantillons, et ce, pour deux populations indépendantes [337]. Cette procédure est ensuite appliquée à 11 populations variées (commerciales *vs.* expérimentales / pontes *vs.* chairs) afin d'estimer le nombre de SNP et génotypes détectables. Nous avons ensuite *i)* analyser les conséquences prédites des variations dans les régions codantes ; *ii)* étudier le potentiel du RNAseq pour les analyses ASE ; *iii)* explorer la diversité génétique entre les 11 populations.

2.1.2. Résultats et discussion

Le génome de poule selon galgal5 et avec une annotation enrichie en lncRNA est composé à parts égales de séquences intergéniques (50 %) et géniques (50 %), avec 43 % d'introns et 7 % d'exons (voir Figure 3A). En utilisant le foie de 15 poules pondeuses, le DNAseq 20X a permis de trouver 7 786 492 SNP bialléliques contre 1 369 740 SNP pour le RNAseq. Moins de SNP sont détectés par RNAseq car seules les régions transcrites sont analysées. Comme attendu (voir Introduction §3.1), du fait d'une pression de sélection inégale [505], les SNP détectés par DNAseq 20X se trouvent principalement dans les régions non-codantes avec 46 % des SNP dans les régions intergéniques et 52 % dans les introns et 2 % dans les exons. Pour les SNP détectés par RNAseq, la majorité se révèlent également être dans les introns (61 %) et régions intergéniques (29 %). La forte part de SNP introniques peut s'expliquer par la présence de transcrits non matures qui ont par définition une expression faible mais suffisante pour la détection de variant, ainsi qu'une pression de sélection moindre dans les introns. Les SNP des régions "intergéniques" apparaissent probablement du fait de régions non annotées correspondant à des UTR voire à des gènes entiers. La répartition des SNP exoniques entre 3'UTR (32 %), 5'UTR (7 %) et CDS (61 %) est similaire entre DNAseq 20X et RNAseq, mais

différente de la proportion génomique de ces régions (20 %, 5 %, 75%), soulignant une pression de sélection moindre dans les 3'UTR.

Afin de pouvoir comparer les méthodes sur un même pied d'égalité, seuls les SNP dans les exons exprimés des gènes exprimées ont été analysés. À cette échelle, 85,2 % des 234 500 SNP trouvés par DNAseq 20X, considérés comme la « vérité », l'ont également été par RNAseq (*i.e.,* sensibilité du RNAseq). Dans une seconde population composée de foie de huit poulets de chairs et donc avec un nombre plus réduit d'échantillons, cette sensibilité était de 65,7 % (voir Figure 3B). Indépendamment de la population étudiée, environ 91 % des SNP détectés par RNAseq l'étaient aussi par le DNAseq 20X, montrant ainsi une bonne précision du RNAseq. Ces résultats sont cohérents avec ceux présentés dans l'étude de Guo et al. [506] qui a identifié environ 85 % de concordance entre RNAseq et séquençage d'exome.
Concernant les 9,4 % de SNP spécifiques au RNAseq, différents facteurs pouvant expliquer leur détection ont été analysés (voir Figure 3C) :

- 46,6 % de ces SNP appartenaient à des « *clusters* » – définis comme un ensemble de 3 SNP ou plus dans une fenêtre de 35 pb – contre 40,0 % des SNP DNAseq spécifiques. Ce filtre qui est conseillé par GATK pour le RNAseq n'est pas recommandé pour le DNAseq. Ce filtre enlevait 39 783 SNP détectés par les deux méthodes engendrant un gain limité de précision (90,6 % à 93,5 %) au regard d'une importante perte de sensibilité (85 % à 68 %) et n'a donc pas été conservé ;
- 5,09 % des SNP RNAseq spécifiques étaient situés au niveau de jonctions d'épissages contre 3,55 % pour les SNP DNAseq spécifiques. Cette différence significative s'explique par la difficulté d'alignement des reads épissés en RNAseq ;
- Dans les régions composées de 5 nucléotides répétées, les proportions de SNP pour le DNAseq et RNAseq atteignaient les 3,5 % et n'étaient pas significativement différentes ;
- En moyenne, 5,5 SNP étaient identifiés dans les 3'UTR pour le RNAseq contre 2,5 pour le DNAseq. Cette différence significative peut être expliqué par la dégradation en 3' des ARNm matures [507].
- L'édition d'ARN pourrait aussi expliquer certains SNP RNAseq spécifiques. Cependant, la littérature suggère que cela constitue un processus rare.
- Les SNP détectés par une seule méthode étaient soutenus par significativement moins de reads que ceux détectés par les deux méthodes.

En utilisant à la fois des échantillons de foie, mais également de sang et d'hypothalamus prélevés sur les mêmes 15 animaux de la première population, l'effet du nombre de tissus sur

la détection de SNP a été étudié (voir Figure 3D). 1 369 740 SNP ont été détectés dans le foie, 1 481 627 dans le sang et 1 511 909 dans l'hypothalamus, alors que 16 814, 16 346 et 19 733 gènes étaient exprimés respectivement dans ces trois tissus. Comme attendu, la combinaison de tissus augmente le nombre de SNP détectés, en relation avec le nombre de gènes exprimés. Pour des études dans lesquelles les RNAseq de différents tissus sont disponibles, il est ainsi conseillé de regrouper les fichiers de séquençage (*.fastq*) par animal avant alignement pour augmenter la puissance et la fiabilité de détection des SNP. Pour les SNP détectés dans plus d'un tissu, la concordance des génotypes entre tissus était très élevée, atteignant 98,9% voire 99,5% et 99,9% pour un filtre imposant 5 et 10 reads respectivement.

Concernant la recherche de génotype (GT) à l'échelle de l'individu, d'autres filtres doivent être considérés. En effet, le RNAseq contrairement au DNAseq 20X ne couvre pas le génome de façon homogène et possède des profondeurs (*depth* notée DP) variables pouvant même être nulles et ainsi le GT n'est pas renseigné (voir Figure 1). Suite à ce constat, la précision du RNAseq pour la détection de GT a été évaluée en considérant le DNAseq 20X comme la vérité. Cette détection est dépendante du nombre d'individus avec un GT renseigné dans la population (*call-rate* noté CR) et de la DP qui sont interdépendants. Ainsi, sans application de filtre, une concordance d'environ 90 % est observée (voir Figure 4). Cette concordance atteint 95 % pour un CR ≥ 20% et une DP ≥ 5, et 97 % pour un CR ≥ 20% et une DP ≥ 10. Il a ainsi été proposé de sélectionner les SNP ayant une DP ≥ 5 *reads* pour au moins 20 % des individus ainsi qu'un CR ≥ 50%, afin de maintenir un nombre d'individus avec GT conséquent. L'application de ces filtres assure une concordance des GT d'au moins 95 % entre le RNAseq et le DNAseq 20X même si la plupart des SNP affiche une concordance > 97%.

Les travaux précédents ont été appliqués pour la détection des SNP et des génotypes dans 11 populations de poules (voir Table 1). Pour les RNAseq de foie, entre 1,1M et 3,8M de SNP ont été détectés par population. Avec tous les tissus disponibles (1 à 5 selon la population), plus de 1,7 à 5,5M de SNP ont été trouvés. Au total, en considérant toutes les populations et tous les tissus, 9,5M de SNP (union) ont été identifiés et 241 960 SNP (intersection) sont présents dans toutes les populations. A titre de comparaison, l'union contient 23 % (2,175M) de SNP non rapportés dans la base dbSNP de Ensembl v94 [508] qui compte 23,8 M de SNP.

Concernant la puce HD 600K de poule [411] largement utilisée pour les études GWAS, seul 5,1 % de ces SNP étaient présents chez l'ensemble de nos individus.
À l'échelle des génotypes et en appliquant les filtres susmentionnés, entre 0,4M et 1,7M de SNP ont été détectés avec tous les tissus, soit 37 % des SNP initiaux, ce qui représente une union de 3,3M de SNP et une intersection de 73 223 SNP. À l'échelle d'un seul tissu, ici le foie, l'union et l'intersection sont du même ordre de grandeur (resp. 1,7M et 67 341 SNP). Après sélection des SNP avec une MAF ≥ 10%, l'union multi-tissus atteint 2,2M et 1,3M pour le foie seul, et environ 2 000 SNP pour l'intersection.

L'impact fonctionnel sur la protéine des 9,5M de SNP détectés dans au moins une population ont été prédits par VEP résultant en 33 304 412 conséquences (voir Figure 5). En cohérence avec les résultats évoqués en début de partie, la grande majorité affectait des régions non-codantes, cependant, parmi les 472 319 SNP affectant une région codante, 63 % étaient synonymes (*synonymous variant*) et 28 % étaient faux sens (*missense variant*). Les 25 344 conséquences délétères dites sévères selon gnomAD [494] ont alors été identifiés avec 590 *stop gained*, 8 126 modifications de sites d'épissage et 16 307 *missense variant* délétères. Parmi elles, 22 % ne présentaient pas de GT ALT/ALT parmi les 382 individus étudiés et 31 % avaient une fréquence ALT/ALT ≤ 5%. Ces résultats suggèrent un rôle important pour ces gènes avec variants à impact sévère prédits. Deux exemples de *missense variant* délétères (SIFT = 0) sont ainsi présentés (voir Figure 6) dont *i)* XBP1, observé à l'état hétérozygote dans deux populations, mais jamais homozygote ALT/ALT et *ii)* SERGEF, présent dans deux populations uniquement à l'état hétérozygote. Ces données constituent donc une ressource complémentaire à la base de données dbSNP de Ensembl pour explorer les fréquences génotypiques et alléliques de variants dans différentes populations.

Puisque le RNAseq permet de détecter, dans les régions exprimées, de nombreux SNP fiables et potentiellement hétérozygote, il a été mis à profit pour explorer le déséquilibre d'expression entre les deux chromosomes parentaux (ASE – voir Figure 7). Nous avons alors comparé l'impact de ne considérer que les SNP des exons ou en incluant les introns. En moyenne, le nombre de gènes avec au moins un SNP hétérozygote est similaire entre transcrits matures et transcrits immatures. Avec les ARN matures exoniques seulement, en moyenne 17 à 28 SNP par gène étaient identifiés, un nombre suffisant pour étudier l'ASE le

long des gènes. Ce nombre est plus élevé pour les lncRNA que pour les PCG (22-28 versus 15-17), probablement à cause d'une pression de sélection moindre. Après application des filtres sur le génotype et une MAF ≥ 10%, 81 % des PCG et 68 % des lncRNA (TPM ≥ 1) apparaissent analysables par ASE. Ce taux décroit respectivement jusqu'à 72 % et 56 % en appliquant un filtre supplémentaire sur l'hétérozygotie (≥ 25%). Nous avons ensuite estimé le nombre de gènes *cis*-régulés dans un tissu par une analyse ASE en utilisant deux populations (RpRm : n = 15 et FLLL : n = 8) et le tissu foie. En utilisant les SNP exoniques et introniques et avec au moins 10 *reads* sur un haplotype, 29 % des PCG/lncRNA exprimés (TPM ≥ 1) étaient *cis*-régulés (34 % pour RpRm et 23 % pour FLLL). Parmi les gènes *cis*-régulés, 50 % des PCG et 37 % des lncRNA étaient partagés entre les deux populations. Ces résultats sont cohérents avec la littérature, qui rapporte 15 % de gènes *cis*-régulés dans le foie embryonnaire de la poule [509] et 26 % dans le foie humain [389].

Nous avons ensuite utilisé les fréquences génotypiques des SNP détectés par RNAseq pour établir des liens génétiques entre les populations (voir Figure 8). La classification obtenue avec l'intersection des SNP génotypés des 10 populations pour le tissu foie (67 341 SNP) est cohérente avec l'histoire connue des populations de poule analysées. Elle sépare clairement la lignée ancestrale *Red Jungle Fowl*, les lignées de chair (*broilers*), les lignées pondeuses à œufs bruns, et les lignées pondeuses à œufs blancs. Les sous-groupes en fonction des orientations commerciales ou expérimentales sont aussi observés. Pour finir, les sous-populations sélectionnées de façon divergente pour un trait spécifique sont également bien distinguées.

2.1.3. Matériels et démarches

Pour la comparaison des SNP détectés par RNAseq et DNAseq, les données de séquençage ont été obtenues à partir des mêmes échantillons de foie collectés sur les mêmes oiseaux de deux populations indépendantes de poule. La première population était composée de 15 individus d'une lignée de ponte expérimentale divergente pour l'efficacité alimentaire (RpRm) tandis que la seconde incluait 8 poules issues d'une lignée de poulet de chair expérimentale (FLLL) divergente pour le gras abdominal [510, 511]. Par la suite, des données RNAseq poly-A disponibles dans les archives publiques ENA et SRA de 11 populations ont été utilisées. Le panel est constitué de :

- 1 population de Red jungle fowl (race ancestrale) ;
- 3 populations de poulets de chair – une expérimentale et deux commerciales ;
- 6 populations de poules pondeuses incluant deux lignées commerciales, deux lignées expérimentales à œufs marrons, une lignée de type Leghorn et une lignée rustique égyptienne ;
- 1 population expérimentale issue d'un croisement entre 2 lignées elles-mêmes expérimentales [345].

La détection des variants à partir des données RNAseq a été effectuée via un *pipeline* snakemake (voir Figure 2) [512]. Brièvement, après contrôle qualité, les séquences ont été alignées (2-pass) à l'aide de STAR [306] sur le génome de référence galgal5 et avec une annotation du génome (*.gtf*) enrichie en lncRNA [124]. Les *reads* s'alignant de façon unique ont ensuite été traités selon « les meilleures pratiques pour les données RNAseq » telles qu'indiquées par GATK. Les variants ont été détectés par échantillon à l'aide de la fonction HaplotypeCaller de GATK. GenotypeGVCFs a ensuite été utilisé pour obtenir les génotypages par tissue (un fichier.*vcf* par tissu). Les SNP bialléliques ont été extraits et filtrés selon deux des trois critères conseillés par GATK à savoir : « *QualByDepth* (QD) < 2 » qui mesure par variant la qualité normalisée par la profondeur de l'allèle ; et « *FisherStrand* (FS) > 30 » qui permet d'identifier les cas anormaux où un brin est favorisé par rapport à un autre. La non-application du filtre « *SnpCluster* » identifiant les SNP regroupés par 3 ou plus dans une fenêtre de 35 pb est par ailleurs discutée dans la partie « Résultats ». Pour l'extraction des génotypes et en accord avec les critères établis dans la section « Résultats et Discussions », seuls les SNP ayant des génotypes renseignés dans au moins 50 % des individus (CR ≥ 50%) et supportés par au moins 5 *reads* pour au moins 20 % des individus ((5.reads.DP) genotype CR ≥ 20%) ont été

sélectionnés. Les fréquences génotypiques et alléliques ont ensuite été calculées à l'échelle des populations, permettant de travailler sur des SNP sélectionnés sur la fréquence de l'allèle mineur (MAF).

Pour le DNAseq, l'alignement des *reads* a été fait avec l'algorithme BWA-MEM [10] sur le génome de référence galgal5. La détection des variants a été réalisée par échantillon avec la fonction HaplotypeCaller de GATK [513–515] ainsi qu'à l'échelle de la population avec GenotypeGVCFs. Les SNP bialléliques ont été extraits et filtrés selon les six filtres recommandés par GATK :

- « FS > 60.0 » ;
- « QD < 2.0 » ;
- « StrandOddsRatio (SOR) > 3.0 » qui a la même utilité que FS mais qui a été mis en place car FS a tendance à pénaliser les variants en extrémité d'exons ;
- « RMSMappingQuality (MQ) < 40.0 » qui prend en compte la variation des qualités des *reads* ;
- « MappingQualityRankSumTest (MQRankSum) < -12.5 » comparant les qualités des *reads* soutenant l'allèle de référence et l'allèle alternatif ;
- « ReadPosRankSumTest (ReadPosRankSum) < -8.0 » qui détermine si les positions des allèles de référence et des allèles alternatifs sont différentes dans les *reads*.

En utilisant l'annotation du génome enrichie en lncRNA et selon le génome de référence galgal5, l'expression des gènes a été quantifiée avec RSEM [355] tandis que l'expression des exons a été réalisée à l'aide de FeatureCount [516]. Un indice RpKb (*Reads par Kilobase*), défini comme le nombre moyen de *reads* mappés sur l'exon divisé par sa longueur en kb a été mis en place pour définir un seuil d'expression exonique. Les valeurs de RpKb ont été comparées à l'expression de loci artificiellement positionnés de manière aléatoire dans le génome pour représenter le bruit de fond et un seuil d'expression de $\log_{10}(RpKb+1) \geq 0,5$ a ainsi été défini pour les exons.

L'annotation fonctionnelle des 9 496 283 SNP identifiés a été réalisée avec VEP *(Variant Effect Predictor*) avec l'option « --everything » pour obtenir les scores SIFT [517, 518].
L'analyse de classification hiérarchique a été effectué avec la fonction "snpgdsHCluster" du package SNPRelate v1.8.0 [519] sur un ensemble de 67 341 SNP obtenus à partir des données

de RNAseq de foie de l'ensemble des populations et répondant aux critères de génotypes fiables évoqués ci-dessus pour chaque population.

Avant de quantifier l'expression allèle-spécifique (ASE), les séquences doivent être alignées sur une version masquée du génome pour éviter de favoriser les allèles de référence. Au niveau de la population, les SNP polymorphes et bialléliques filtrés avec GATK, ont servi à masquer le génome de référence et les données RNAseq ont ainsi pu être alignées avec STAR (2-pass). phASER Gene AE [520] a été utilisé pour détecter l'ASE dans les échantillons de foie. Rapidement, phASER détecte les haplotypes, compte les *reads* associés à chacun et en sélectionne un par gène pour étudier l'ASE. Seuls les gènes avec au moins 10 *reads* par haplotype ont été conservés. Un test binomial a permis de détecter les déséquilibres de nombre de *reads* entre haplotypes parentaux. Les p-values ont été corrigées avec la méthode de Benjamini-Hochberg [456]. Un gène a été considéré ASE s'il présentait un déséquilibre significatif entre les deux haplotypes dans au moins deux échantillons.

2.1.4. Valorisations associées

Ces travaux ont fait l'objet :

- d'une publication : Jehl F*, **Degalez F***, Bernard M*, Lecerf F, Lagoutte L, Désert C, Coulée M, Bouchez O, Leroux S, Abasht B, Tixier-Boichard M, Bed'hom B, Burlot T, Gourichon D, Bardou P, Acloque H, Foissac S, Djebali S, Giuffra E, Zerjal T, Pitel F, Klopp C and Lagarrigue S (2021). RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Frontiers in Genetics. doi: 10.3389/fgene.2021.655707. ***Cet article est reproduit ci-après ;***

- d'une présentation orale : Jehl F*, **Degalez F***, Bernard M*, Lecerf F, Lagoutte L, Désert C, Coulée M, Bouchez O, Leroux S, Abasht B, Tixier-Boichard M, Bed'hom B, Burlot T, Gourichon D, Bardou P, Acloque H, Foissac S, Djebali S, Giuffra E, Zerjal T, Pitel F, Klopp C and Lagarrigue S (2022). RNA-seq data for detecting reliable SNPs & genotypes in livestock species: interest for coding variant characterization and cis-regulation analysis by allele-specific expression. Communication faite par Sandrine Lagarrigue dans la session spécialisée "Molecular Genetics" au "World's Poultry Congress" (WPC), Paris, France ;

- d'un « e-poster » (présenté à distance) : Jehl F, **Degalez F**, Bernard M, Lecerf F, Coulee M, Zerjal T, Pitel F, Klopp C, Lagarrigue S. (2020). Genomic SNP detection by RNA-seq: lessons from multi-tissue & multi-population data analysis in chickens. Communication réalisée aux "Open Day of Computational Biology and Mathematics" (JOBIM), Montpellier, France.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fgene.2021.655707

RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and *Cis*-Regulation Analysis by Allele-Specific Expression in Livestock Species

Frédéric Jehl[1†], Fabien Degalez[1†], Maria Bernard[2,3†], Frédéric Lecerf[1], Laetitia Lagoutte[1], Colette Désert[1], Manon Coulée[1], Olivier Bouchez[4], Sophie Leroux[5], Behnam Abasht[6], Michèle Tixier-Boichard[3], Bertrand Bed'hom[3], Thierry Burlot[7], David Gourichon[8], Philippe Bardou[2], Hervé Acloque[3], Sylvain Foissac[5], Sarah Djebali[5], Elisabetta Giuffra[3], Tatiana Zerjal[3], Frédérique Pitel[5], Christophe Klopp[2‡] and Sandrine Lagarrigue[1*‡]*

[1] INRAE, INSTITUT AGRO, PEGASE UMR 1348, Saint-Gilles, France, [2] INRAE, SIGENAE, Genotoul Bioinfo MIAT, Castanet-Tolosan, France, [3] INRAE, AgroParisTech, Université Paris-Saclay, GABI UMR 1313, Jouy-en-Josas, France, [4] INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France, [5] INRAE, INPT, ENVT, Université de Toulouse, GenPhySE UMR 1388, Castanet-Tolosan, France, [6] Department of Animal and Food Sciences, University of Delaware, Newark, DE, United States, [7] NOVOGEN, Maugueérand, Le Foeil, France, [8] INRAE, PEAT UE, Nouzilly, France

In addition to their common usages to study gene expression, RNA-seq data accumulated over the last 10 years are a yet-unexploited resource of SNPs in numerous individuals from different populations. SNP detection by RNA-seq is particularly interesting for livestock species since whole genome sequencing is expensive and exome sequencing tools are unavailable. These SNPs detected in expressed regions can be used to characterize variants affecting protein functions, and to study *cis*-regulated genes by analyzing allele-specific expression (ASE) in the tissue of interest. However, gene expression can be highly variable, and filters for SNP detection using the popular GATK toolkit are not yet standardized, making SNP detection and genotype calling by RNA-seq a challenging endeavor. We compared SNP calling results using GATK suggested filters, on two chicken populations for which both RNA-seq and DNA-seq data were available for the same samples of the same tissue. We showed, in expressed regions, a RNA-seq precision of 91% (SNPs detected by RNA-seq and shared by DNA-seq) and we characterized the remaining 9% of SNPs. We then studied the genotype (GT) obtained by RNA-seq and the impact of two factors (GT call-rate and read number per GT) on the concordance of GT with DNA-seq; we proposed thresholds for them leading to a 95% concordance. Applying these thresholds to 767 multi-tissue RNA-seq of 382 birds of 11 chicken populations, we found 9.5 M SNPs in total, of which ~550,000 SNPs per tissue and population with a reliable GT (call rate ≥ 50%) and among them, ~340,000 with a MAF ≥ 10%. We showed that such RNA-seq data from one tissue can be used to (*i*) detect SNPs with a strong predicted impact on

OPEN ACCESS

Edited by:
James Reecy,
Iowa State University, United States

Reviewed by:
Stephen J. Bush,
University of Oxford, United Kingdom
Melissa Susan Monson,
Iowa State University, United States

***Correspondence:**
Christophe Klopp
christophe.klopp@inrae.fr
Sandrine Lagarrigue
sandrine.lagarrigue@agrocampus-ouest.fr

†These authors share first authorship

‡These authors share last authorship

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *19 January 2021*
Accepted: *01 June 2021*
Published: *28 June 2021*

Citation:
Jehl F, Degalez F, Bernard M, Lecerf F, Lagoutte L, Désert C, Coulée M, Bouchez O, Leroux S, Abasht B, Tixier-Boichard M, Bed'hom B, Burlot T, Gourichon D, Bardou P, Acloque H, Foissac S, Djebali S, Giuffra E, Zerjal T, Pitel F, Klopp C and Lagarrigue S (2021) RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Front. Genet. 12:655707. doi: 10.3389/fgene.2021.655707

proteins, despite their scarcity in each population (16,307 SIFT deleterious missenses and 590 stop-gained), (*ii*) study, on a large scale, *cis*-regulations of gene expression, with ∼81% of protein-coding and 68% of long non-coding genes (TPM ≥ 1) that can be analyzed for ASE, and with ∼29% of them that were *cis*-regulated, and (*iii*) analyze population genetic using such SNPs located in expressed regions. This work shows that RNA-seq data can be used with good confidence to detect SNPs and associated GT within various populations and used them for different analyses as GTEx studies.

Keywords: RNA-seq, SNP calling, genotype calling, SNP annotation, allele-specific expression, livestock, chicken

INTRODUCTION

RNA-seq is currently the method of choice to study transcriptome expression in replacement of gene chips (Mortazavi et al., 2008). This technology is commonly used to study gene expression patterns in a variety of organisms including plant, animal or human groups to better understand the genetic mechanisms intervening in the determinism of phenotypes (Gondret et al., 2017), diseases (Savary et al., 2020) or response to environmental changes (Jehl et al., 2019) among others. The RNA-seq has other more specific applications taking advantage of its sequencing step. For example RNA-seq allows transcript and gene modeling as shown by long non-coding atlas reported in different species (Derrien et al., 2012; Jehl et al., 2020). It also allows to combine SNP information, at the RNA level with gene expression to study the variation which affects gene-expression levels: it is a powerful technology to identify such expression quantitative trait locus (eQTL) either through GWAS mapping (if the individual number is sufficient) or through allele-specific expression (ASE) analysis as shown by growing number of studies on a variety of species since the beginning of the RNA-seq technology in the 2010s (Montgomery et al., 2010; Pickrell et al., 2010; Battle et al., 2013; Lagarrigue et al., 2013b; Chamberlain et al., 2015; Deelen et al., 2015; The GTEx Consortium, 2020), among them the famous studies from the human GTEx consortium (The GTEx Consortium, 2020). Finally RNA-seq allows RNA editing analysis, a phenomenon resulting in nucleotide changes observed at RNA level, occurring after its transcription from DNA level (Kleinman et al., 2012). In these two last applications, RNA-seq is in general combined with DNA-seq used for genotyping individuals. However, RNA-seq can also detect genomic variations in expressed regions like DNA-seq, as described by Piskol et al. (2013). It is particularly interesting in non-model species (wild or domesticated, for example livestock species) in which no exome capturing tools have been developed as an alternative to DNA-seq data, which remains costly to generate and store. In this context, RNA-seq presents several advantages compared to the DNA-seq. First, the number of RNA-seq data sets publicly available is much higher than the number of DNA-seq data sets, for many species (chicken, pig, cow, and other non-model species) since these data have accumulated over the past several years and continue to accumulate in different populations and within populations. Moreover, within populations, different conditions are studied, increasing the number of studied animals, allowing to better detect, in a given population, variants with low frequencies. Second, RNA-seq data allows studying coding region variations that have potential functional impacts. Some of these SNPs can induce a loss of the protein function. These loss-of-function variants are extensively studied because of their possible contribution to phenotypes (Genome Aggregation Database Consortium et al., 2020). In addition, they represent a powerful source of information to understand gene functions (Genome Aggregation Database Consortium et al., 2020). However, these loss of function SNPs are rather rare because purged by negative selection in natural populations but can be detected with a certain number of samples. In well-known model-species or human, these coding region variants are accessible using whole exome sequencing (WES), as shown by the recent work of the Genome Aggregation Database (gnomAD) (Lek et al., 2016). This consortium analyzed 125,748 human exomes (and much fewer whole genomes: 15,708) from public sources and identified 443,769 high-confidence predicted loss-of-function variants, defined in the work of gnomAD as being either gain of stop (non-sense variants), frameshift or splice site variants. For non-model species such as livestock species, for which the WES method is usually not available, RNA-seq can thus fulfill the same objective, with a similar advantage that is, producing a smaller data volume, thus facilitating data storage and decreasing costs (Battle et al., 2013). Third, RNA-seq data provides expression levels of loci harboring SNPs, allowing to study allele-specific expression as we previously mentioned, and hence, to study *cis*-regulation on a large scale, in multiple tissues and multiple populations. Fourth, the transcribed regions are well spread over the genome and much more numerous than previously thought. Thousands of novel long non-coding genes exist across the genome, as highlighted by the ENCODE project (Derrien et al., 2012). RNA-seq data can therefore provide sets of numerous and well distributed SNPs throughout the genome. Finally, these data could be used to study population genetic diversity from a different point of view compared to the SNP chips, by offering various sets of SNPs with more or less severe functional impacts and not neutral SNPs.

Despite the aforementioned advantages RNA-seq is not yet often used for SNP detection in coding regions. Indeed, SNP detection and genotype calling by RNA-seq present three main challenges. First, the transcriptome is composed of mature transcripts (i.e., spliced), making mapping of RNA-seq reads that overlap exon-exon junctions, more difficult, compared to DNA-seq read alignment (Pan et al., 2008). However, RNA-seq

mapping methods seem to be well mastered in recent years, even though it is important to remain cautious for SNPs detected close to exon-exon junctions (Peng et al., 2012; Lagarrigue et al., 2013b). Second, RNA editing, by definition, could represent a strong limitation for SNP detection by RNA-seq, mainly because it introduces variations at the RNA level, which are absent at the DNA level. Nevertheless, as we will discuss later, RNA editing has such features that it only slightly impedes reliable RNA-seq based variation detection in standard conditions. Third, genes exhibit highly variable expression levels, leading to the read depths ranging from a few reads to millions of reads, contrarily to the DNA-seq which offers a rather homogeneous read depth across the genome (see **Figures 1A,B**). Indeed, coding and non-coding transcripts can be expressed at vastly different levels, ranging from few copies to millions of copies per cell, in different cell types and developmental or physiological stages. Moreover, the transcriptome is also composed of a small portion of immature under processing transcripts (composed of exons and introns), less supported by reads but enriched in introns that are more variable in sequence compared to exons (Sims et al., 2014). In summary, these variations in read depth from one gene to another, and within a gene (between introns and exons) constitute a major challenge for SNP detection (see **Figures 1A,B**, left), and more importantly, for individual genotype calling (see **Figures 1A,B**, right). Indeed, reliable SNP detection at the population level benefits from the information accumulation born by the reads across individuals, in contrast to genotype calling. This last point might explain why only few studies have used RNA-seq data for variant detection and genotype calling since the first publications. Consequently, neither the number of SNPs that could be detected using RNA-seq, nor the percentage of individuals with a given genotype (a prerequisite for computing allelic frequencies), are known. To our best knowledge, since Piskol et al. (2013), less than a dozen studies were focused on large-scale SNP detection tools from RNA-seq data (Quinn et al., 2013; Tang et al., 2014; Wang et al., 2014; Wolfien et al., 2016; Oikkonen and Lise, 2017; Cornwell et al., 2018; Adetunji et al., 2019). The reference tools for read mapping and variant detection have been evolving very rapidly, and these studies have tested different tools, and among them, only Adetunji et al., 2019 (Adetunji et al., 2019) used the most recent tools proposed by ENCODE for RNA-seq data, i.e., STAR (Dobin et al., 2013) for read mapping and GATK (Van der Auwera et al., 2013) for variant detection. Three of the above-mentioned studies were interested in determining the concordance of SNP and genotype detection between RNA-seq and DNA-seq, the latter being the gold standard for SNP detection. However, these studies used only few samples (from 1 to 4) and had not at their disposal both RNA-seq and DNA-seq data on the same tissues of the same individuals.

In this context, this work aims at detecting SNPs from RNA-seq data in chicken. The first goal was to set up a procedure allowing SNP detection and genotype (GT) calling from RNA-seq data using reference tools (STAR for read mapping and GATK for SNP detection). We tested the SNP reliability according to three filters suggested by the GATK team and compared the detected SNPs with those obtained using DNA-seq data. This comparison was performed in two independent chicken populations for which RNA-seq and DNA-seq data were available on the same biological samples (i.e., the same tissue of the same

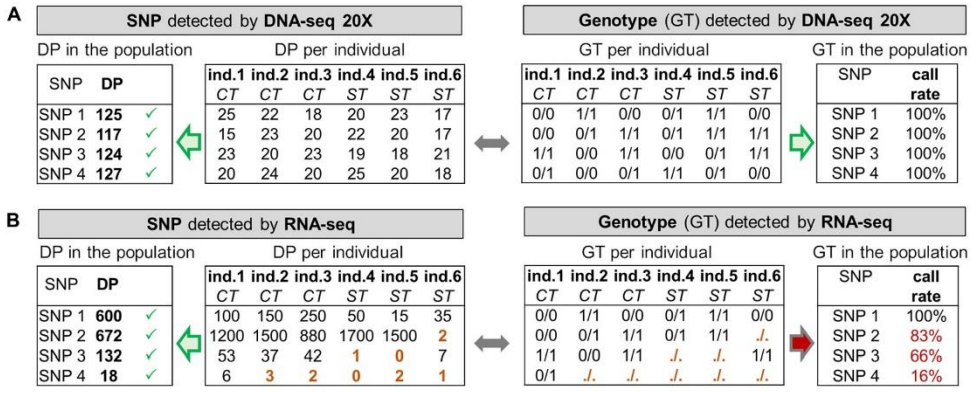


FIGURE 1 | Toy example with simulated data illustrating the need for read depth (DP) filters in RNA-seq and differences with DNA-seq. **(A)** DNA-seq data offers a globally homogeneous genome coverage (20X in our case), all SNPs are therefore detected by GATK at the individual level with a DP of 20 reads on average (*"DP per individual"*), and at the population level with a DP of 6 × 20 = 120 reads on average (*"DP in the population* resulting from the addition of *"DP per individual"*). All genotypes (GT) can therefore be computed at the individual level (*"GT per individual"*), resulting in a genotype call rate of 100% for every SNP (*"GT in the population"*). **(B)** RNA-seq data offers a heterogeneous coverage of the genome depending on the expression of the genes harboring the SNPs. *At the population level,* 4 SNPs having a sufficiently high DP are detected by GATK. At the individual level, SNP 1 shows good read coverage across all samples whereas SNP 3 is on a gene that has a lower expression, in particular in the stress (ST) condition compared to the control (CT). SNP 4 is on an overall very lowly expressed gene. *In terms of genotype (GT) per individual*, some cannot be provided by GATK (noted "./.") because of a too low DP (i.e., 5 reads, see brown GT and DP) and are not considered for the GT call rate. For SNP 3, most of the individuals from the ST condition have no GT and for SNP 4, only one GT is called whereas in both case the SNP is detected at the population-level. "GT in the population" provides for each SNP their call-rate for the genotypes (CR): SNP 1 has 100% of the samples with a GT whereas SNP 4 has 16% and cannot be used to compute meaningful genotype frequencies.

individuals). In this paper, the workflow was used at the tissue level to provide results for RNA-seq experimental settings with only one analyzed tissue which represent a quite common case. This, however, corresponds to the least favorable case compared to multi-tissue experimental projects, since it does not allow cumulating the sequences from tissues per individual. We then analyzed the effects on the number of detected SNPs by this workflow performed at the tissue level when using additional tissues of a same population.

Because a large proportion of SNPs detected by RNA-seq was reliable, we further applied this procedure to 11 different chicken populations: a population derived from the wild Red Jungle Fowl population, an Egyptian Fayoumi population, six commercial and experimental laying hen populations and three commercial and experimental broiler populations. Our three goals were to (*i*) provide an estimation of the number of SNPs and GT that can be detected using RNA-seq data per tissue and population, (*ii*) present an overview of the predicted consequences of the SNPs located in coding regions, in particular, the number of high-confidence predicted loss-of-function variants, as defined in the work of gnomAD, and finally (*iii*) give an overview of the potential of RNA-seq for allele-specific expression (ASE) analysis by estimating the number of genes that could be analyzed for ASE with the number of SNPs detected per gene. We then identified the *cis*-regulated genes in the liver of 2 of the 11 populations using the phASER tool (Castel et al., 2016) and the proportion of cis-regulated hepatic genes shared by the two populations. Finally, we illustrated the possibility of using RNA-seq data to explore genetic diversity between populations using different hepatic RNA-seq SNP sets with variable percentage of severe predicted protein consequence.

MATERIALS AND METHODS

RNA-Seq and DNA-Seq Data

Raw data of both DNA-seq and/or RNA-seq are available on the ENA and SRA archives under accession numbers: PRJEB28745 (RpRm DNA-seq and RNA-seq, Novo1 and Novo2, RNA-seq); PRJEB43829 (FLLL, DNA-seq); PRJNA330615 and PRJNA248570 (FLLL, RNA-seq); PRJEB26695 (red jungle fowl, RNA-seq); PRJEB34341 (Naked neck, RNA-seq); PRJEB34310 (Fayoumi, RNA-seq); PRJEB27455 (FrAg, RNA-seq); PRJEB43662 (Cobb, RNA-seq); PRJNA612882 (HerX, RNA-seq) (Fu et al., 2015). RNA sequencing was conducted on all samples using an Illumina HiSeq (Illumina, California, United States) system, with 2 × 150 bp or 100 bp. Libraries were prepared following Illumina's instructions by purifying poly-A RNAs (TruSeq RNA Sample Prep Kit). Illumina adapters containing indexing tags were added for subsequent identification of samples.

For the comparison of SNPs detected by RNA-seq versus DNA-seq, we used two populations for which both data types were obtained from same liver samples collected on the same birds. The population A was composed of 15 birds from an experimental layer population (RpRm, PRJEB28745) composed of birds diverging for feed efficiency (Rp and Rm) after a 40-year diverging selection (Bordas et al., 1992). The population B was composed of 8 birds from an experimental broiler population (FLLL, PRJNA330615) composed of birds diverging for body fat content (FL and LL) (Roux et al., 2015).

For the rest of the work, we used RNA-seq data from 11 populations (see Additional File 1 for the detail of the number of birds, the tissues and the number of samples): a red jungle fowl population (called RJFh with 36 birds and 3 tissues); 3 broiler populations, the FLLL presented previously but here extended with 32 birds and 2 tissues) and two commercial ones, the Cobb 500 (Cobb Vantress, named Cobb with 48 birds and 2 tissues) and a 3-way cross produced by Heritage Breeders, LLC (named HerX, 23 birds and 1 tissue), 6 layer populations with 2 commercial brown-egg subpopulations from the Novogen company, Novo1 with 32 birds and 1 tissue and Novo2 with 40 birds and 2 tissues, 2 experimental brown-egg populations with the RpRm presented previously but here extended (with 88 birds and 5 tissues) and an experimental dwarf chicken layer line homozygous for the Naked Neck mutation (named LSnu with 16 birds and 2 tissues) and 2 other layer populations with a leghorn breed (FrAg) with 4 birds and 2 tissues) and the Fayoumi (FAyo), an Egyptian breed with 16 birds and 2 tissues; finally an experimental population (Rmx6) issued from crosses between 2 experimental lines (Frésard et al., 2014) with 19 embryos harvested from the same batch at embryonic day 4.5 (stage 26).

RNA-Seq Read Mapping and Variant Detection

For all samples, RNA-seq variants were detected using the snakemake (Koster and Rahmann, 2012) pipeline, available at this reference: (GitLab, 2019). For each population, samples were analyzed by tissue. FASTQ files were trimmed for Illumina adapter using TrimGalore version 0.4.5 (Krueger, 2021). STAR v.2.5.2b (Dobin et al., 2013) was used with default parameters for the read mapping on the Gallus_gallus-5.0 reference genome, after the multi-sample 2-pass mapping procedure, with a GTF file enriched in long non-coding genes [available on http://www.fragencode.org (LNChickenAtlas); Section: Galgal5—Ensembl v94; Genome annotation: LNCextendedEns94.gtf.gz; (Jehl et al., 2020)]. Uniquely mapped reads (selected on a mapping quality score equal to 255) were then post-processed following the GATK best practices for RNA-seq data [duplicates were marked, reads overlapping intron were split and mapping quality score were reassigned, indel were realigned and base were recalibrated thanks to the known variants from Ensembl v94's dbSNP (Ensembl, 2018)]. Variant detection was done for each sample using the "HaplotypeCaller" function of GATK (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013) 3.7.0 with option "—stand_call_conf 20.0," "—min_base_quality_score 10" and "—min_mapping_quality_score 20" (which are the defaults values), generating one gVCF file per sample. The "GenotypeGVCFs" function was then used with option "—stand_call_conf 20.0," to jointly genotype all these samples into one VCF per tissue. The VCF file obtained at the end of the pipeline was then used as the input to two other steps,

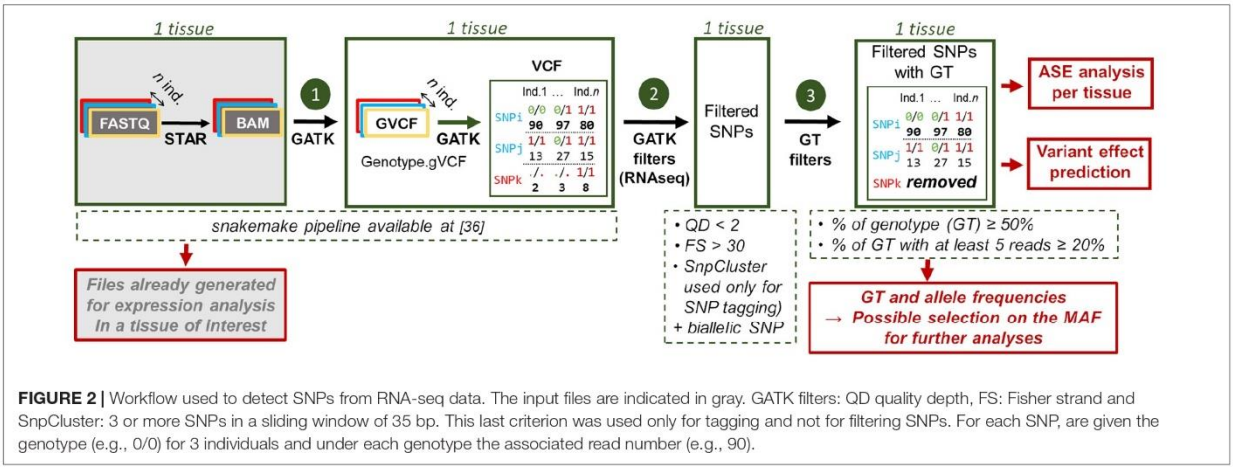


FIGURE 2 | Workflow used to detect SNPs from RNA-seq data. The input files are indicated in gray. GATK filters: QD quality depth, FS: Fisher strand and SnpCluster: 3 or more SNPs in a sliding window of 35 bp. This last criterion was used only for tagging and not for filtering SNPs. For each SNP, are given the genotype (e.g., 0/0) for 3 individuals and under each genotype the associated read number (e.g., 90).

as summarized in **Figure 2**. First, biallelic SNPs were then extracted using the "SelectVariant" function with option "—selectType SNP—restrictAllelesTo BIALLELIC." Variants were also filtered using "VariantFiltration" with two of the three suggested filters, "QD < 2" and "FS > 30," as we discussed in the Results and Discussion section. Finally, we selected the SNPs with genotypes associated with each individual and that met the criteria established in results and Discussion section, i.e. (5.reads.DP) genotype CR ≥ 20% and CR ≥ 50%. Genotype and allele frequencies were then computed, making possible to work on SNPs selected on the minor allele frequency (MAF). These VCF files containing the SNP with their associated genotypes can be used for allele specific expression (ASE) analysis in each tissue of interest.

It is important to note that all previous treatments were conducted in this paper at the tissue level to provide SNP detection results for RNA-seq experimental settings with only one analyzed tissue, which is quite common and corresponds to the least favorable case. This implies that we had one bird's genotype per tissue. For the multi-tissue analysis step of this paper, gVCF files generated per tissue were combined and genotypes were computed from all the tissues information using "CombineGVCFs" and "GenotypeGVCFs" generating per bird as many genotypes as tissues analyzed. Genotype concordance between tissues for a same bird was very high (∼99% of SNPs) and increased with coverage (see result section). Therefore, for the rare cases of discordance, we kept the genotype of the tissue with the highest coverage when they were different. However, outside from this study, for projects in which RNA-seq of different tissues per animal are available when the SNP detection analysis is started, we advise users of our pipeline to define in the first step a sample as a specific individual. This strategy allows to gain power in SNP detection by gathering all BAM tissue files per animal.

DNA-Seq Read Mapping and Variant Detection

DNA-seq read mapping and variant detection were performed using standard tools. The BWA-MEM algorithm (Li, 2013) from BWA-0.7.17 was used with default parameters for the read mapping on the Gallus_gallus-5.0 reference genome (GCA_000002315.3). Variant detection was done for each sample using the "HaplotypeCaller" function of GATK (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013) 3.7.0 with option "-variant_index_type LINEAR," "-variant_index_parameter 128000," "-mmq 30" and "-mbq 10 2," generating one gVCF file per sample. The "CombineGVCFs" and "GenotypeGVCFs" (with "stand_call_conf 20.0" option) functions were then used to combine these gVCF into one VCF per population (one VCF for the 15 RpRm and one VCF for the 8 FLLL). Biallelic SNPs were then extracted using the "SelectVariant" function with option "—selectType SNP—restrictAllelesTo BIALLELIC." Variant were filtered using "VariantFiltration" with all the recommended filters for DNA-seq: "FS > 60.0," "QD < 2.0," "SOR > 3.0," "MQ < 40.0," "MQRankSum < −12.5" and "ReadPosRankSum < −8.0."

Gene and Exon Expression Quantification

Gene expression was quantified with RSEM (Li and Dewey, 2011) v.1.3.0, at the gene-level, using the GTF file LNCextendedEns94.gtf.gz available on http://www.fragencode.org (LNChickenAtlas; section Galgal5) and corresponding to the genes from the Ensembl annotation used as reference, extended with lncRNAs loci available in other public databases (NCBI, NON-CODE, etc.) (Muret et al., 2017). To compute expression at the exon level, we used FeatureCount v1.6.2 (Liao et al., 2014) with options -t "exon" and -g "exon_id." We defined for each exon a metric called RpKb (Read per Kilobase) as the mean number of reads mapped at the exon divided by its length in kilobases. To define an expression threshold, we compared the expression of exons to the expression of a set of randomly selected loci in the genome as done previously in Jehl et al. (2020). The background noise corresponds to the expression of a set of artificial loci randomly distributed across chicken chromosomes 1–33 using the "shuffle" function from the BEDTools suite

v2.29 (Quinlan and Hall, 2010). These artificial loci had the same length distribution as the LNC genes known to be the less expressed compared to PCG and were positioned at a distance of at least 5kb of the closest known transcribed regions. The expression of these randomly selected regions was well below the expression of the exons. We set as an expression threshold for the exons a $\log_{10}(\text{RpKb} + 1)$ value of 0.5, corresponding to the first quartile of expression in both RpRm and FLLL (see Additional File 2).

Variant Functional Predictions

Variant Effect Predictor (VEP) v92 (McLaren et al., 2016) with the GTF file enriched in long non-coding genes ("—gtf") was used for effect prediction of 9,496,283 SNPs. "—everything" and "—total_length" options were applied to respectively, obtain SIFT score predictions and length of cDNA, CDS and proten positions (Ng, 2003; Sim et al., 2012).

Detection of Homopolymers and Exon-Exon Junctions

Regions with 5 or more repeated nucleotides (homopolymers) and regions spanning 5 bp of each extremity of a junction were detected using home-made scripts.

Hierarchical Clustering Analysis

The hierarchical clustering was performed on a set of 67,341 SNPs obtained using liver RNA-seq data from the 10 populations presented in **Table 1** (liver unavailable for Rmx6). This set corresponds to the SNPs common to the 10 populations and passes the GT criteria (see "Results and discussion") for each population. The analysis was produced by using the function "snpgdsHCluster" of the R (R Core Team, 2019) package SNPRelate v1.8.0 (Zheng et al., 2012).

Allele-Specific Expression (ASE) Analysis

Prior to the quantification of allele specific expression, sequences need to be aligned against masked version of the genome to avoid favoring reference alleles. At the population level, polymorphic (allele frequency < 100%) and bi-allelic filtered (GATK—FS and QD criteria) SNP were extracted using the GATK "SelectVariants" tool. These last variants were then used to mask the reference genome using "maskfasta" tool from the BEDTools suite v2.29. Tissue sample sequence were aligned to this masked version of the genome using the multi-sample 2-pass mapping procedure of STAR 2.6. Non-duplicated ("MarkDuplicates" function from GATK 4.1.2, with "READ_NAME_REGEX" set to null) properly paired (if paired sequences) uniquely mapped reads (samtools 1.9 with –f 2 and –q 255 options) were selected. "SplitNCigar" tool from GATK were finally used to split alignment overlapping exon/intron junction and rescaled mapping quality. The phASER tool (Castel et al., 2016) and its downstream tool phASER Gene AE were used to detect ASE among the liver samples of the RpRm and FLLL populations. Briefly, phASER phases, in each sample, SNPs from a user-provided VCF, using the reads from the previously processed BAM file of the sample. This produces a list of haplotypes upon which phASER counts the number of reads associated to each "super-allele." Then, in each sample, phASER Gene AE selects one haplotype per gene, using the genes' boundaries from a user-provided BED file, allowing the study of the gene's ASE using the selected haplotype.

Using base quality of 10, and mapping quality of 20, we provided a VCF containing the SNP that met the criteria established here. After selection of one haplotype per gene using phASER Gene AE, we considered only the genes represented by a haplotype with at least 10 reads associated to at least 1 super-allele. To assess ASE in each sample, we screened for read number imbalance between the super-alleles using a binomial test

TABLE 1 | SNP counts per population retained at each step of the selection.

Population				Total SNP			Selected GT					Selected GT and MAF ≥ 10%				
Pop.	#ind.	#smpl.	#tiss.	Liver[a]	Multi-tiss.[b]	b/a	Liver[c]	Multi-tiss.[d]	d/c	c/a	d/b	Liver[e]	Multi-tiss.[f]	f/e	e/a	f/b
RJFh	36	72	3	1,050,035	2,604,288	2.48	265,750	578,726	2.18	0.25	0.22	152,029	319,268	2.10	0.14	0.12
Cobb	48	96	2	3,771,992	5,464,266	1.45	949,127	1,678,364	1.77	0.25	0.31	558,020	952,445	1.71	0.15	0.17
FLLL	32	64	2	1,729,800	2,033,207	1.18	535,228	1,109,324	2.07	0.31	0.55	368,280	714,523	1.94	0.21	0.35
HerX	23	23	1	1,332,709	1,332,709	1.00	481,314	481,314	1.00	0.36	0.36	307,859	307,859	1.00	0.23	0.23
Novo1	32	32	1	1,459,352	1,459,352	1.00	447,594	447,594	1.00	0.31	0.31	264,804	264,804	1.00	0.18	0.18
Novo2	44	104	2	1,289,199	2,146,975	1.67	390,195	738,109	1.89	0.30	0.34	243,892	449,446	1.84	0.19	0.21
RpRm	112	286	5	1,841,778	4,032,988	2.19	555,928	1,279,458	2.30	0.30	0.32	307,049	631,868	2.06	0.17	0.16
Rmx6	19	19	1	–	2,123,217	–	–	715,822	–	–	0.34	–	483,379	–	–	0.23
FrAg	4	7	2	1,247,253	1,732,440	1.39	784,397	1,055,772	1.35	0.63	0.61	520,277	583,742	1.12	0.42	0.34
Lsnu	16	32	2	1,487,176	2,284,902	1.54	590,399	836,800	1.42	0.40	0.37	384,720	534,938	1.39	0.26	0.23
Fayo	16	32	2	1,320,244	2,033,207	1.54	496,412	698,932	1.41	0.38	0.34	288,464	396,446	1.37	0.22	0.19
Mean				1,652,954	2,477,050	1.54	549,634	874,565	1.64	0.35	0.37	339,539	512,611	1.55	0.22	0.22
Union	382	767		5,490,587	9,496,283		1,685,406	3,276,615				1,255,554	2,243,766			
Intersection				221,374	241,960		67,341	73,223				2,442	1,442			

In columns—Pop., population; #ind., bird number; #smpl., sample number; #tiss., tissue number; Multi-tiss., Multi-tissues. Superscripts are used to show which ratio are presented. Total SNP: SNPs detected at the population level (i.e., with at least one ALT allele); Selected GT: SNPs with at least 50% of genotypes (CR ≥ 50%) and 20% of GT with reads ≥ 5 reads [(5.reads.DP)genotypeCR ≥ 20%, see "Results and discussion"]; Selected GT with minor allele frequency (MAF) ≥ 10%.
In lines—Union: SNPs detected in at least one population; Intersection: SNPs detected in each of the 10 populations (i.e., each population has at least one ALT allele).

(*binom.test* R function) with the null hypothesis that, for a given gene, each super-allele had the same number of associated reads. *P*-values were corrected using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) with a false discovery rate of 0.05. We considered a gene to be ASE if it presented a significant read number imbalance in at least 2 samples.

RESULTS AND DISCUSSION

SNP Detection by RNA-Seq: Genome Location

We compared the repartition of the SNPs detected by DNA-seq and RNA-seq among different genomic regions (**Figure 3A**). The chicken genome is composed at equal parts of intergenic (50%) and genic (50%) sequences, with 43% of introns and 7% of exons. As expected, DNA-seq SNPs were mostly distributed across the non-coding part of the genome (46% in intergenic regions, 52% in introns) and at a lower proportion (2%) in exonic regions. This distribution is expected since coding regions are generally under stronger selection pressure than non-coding regions (Zhao et al., 2003). With RNA-seq (all the samples being systematically treated with DNAse), we expected to find most of the SNPs in exonic regions, which represent the majority of expressed regions. However, the majority of the detected SNPs were located in intronic (61%) and intergenic (29%) regions. Higher SNP counts in intronic regions can be explained by the presence of unspliced transcripts (premature transcripts), very lowly expressed compared to spliced transcripts, but sufficiently to be supported by reads, and by the lower selection pressure on these regions compared to the exons. SNPs located in "intergenic regions" are likely to be located in new genes or in not yet annotated part of genes (particularly 3′UTR and 5′UTR). Within exons, the proportion of SNPs in 3′UTR, 5′UTR and CDS were similar between RNA-seq and DNA-seq (32, 7, 61%), but significantly different from the proportion of these regions in the genome (20, 5, 75%) showing a lower selection pressure in 3′UTR regions than in CDS regions.

SNP Detection by RNA-Seq: Concordance With Those Detected by DNA-Seq

We detected SNPs using either RNA-seq or DNA-seq data obtained from the liver of the same 15 laying hens (see population A in **Figure 3B**, left). We found 7,786,492 biallelic SNPs using the DNA-seq data filtered with the standard criteria of GATK (see section "Materials and Methods") and considered them as reliable. Using the RNA-seq data filtered with some of the filters

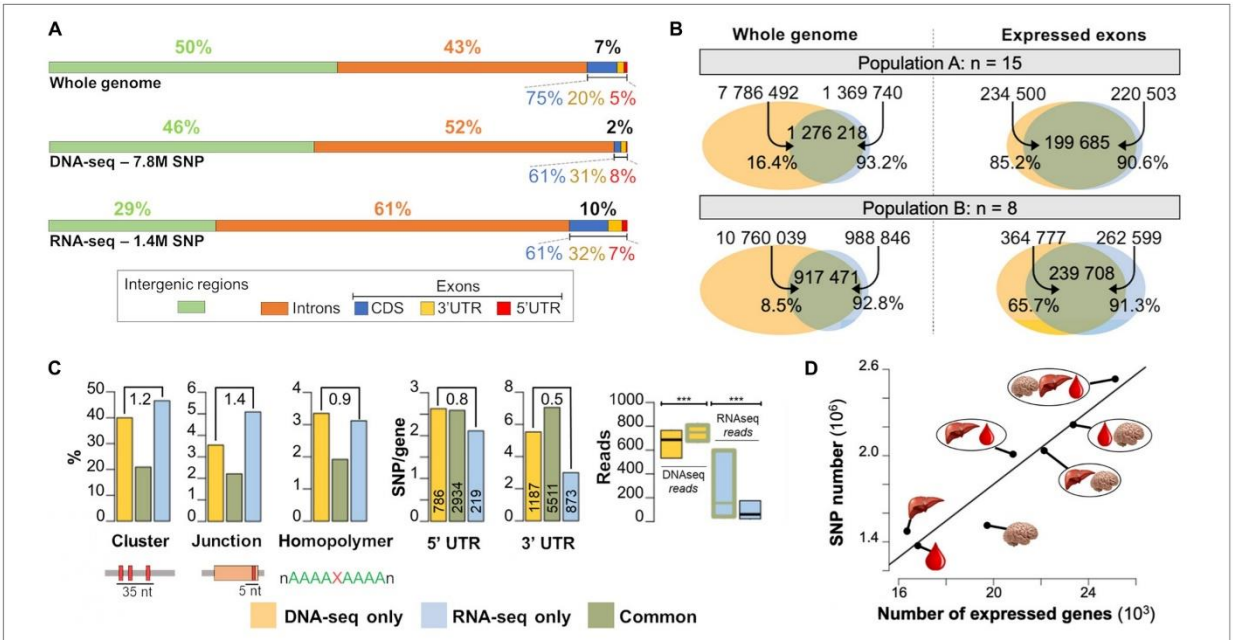


FIGURE 3 | Differences and common features of SNPs detected by RNA-seq and DNA-seq. **(A)** Percentage of the genome comprising each type of feature (top) and the proportion of SNPs detected by DNA-seq (middle) and RNA-seq (bottom) across these genomic features. **(B)** Number of SNPs detected by DNA-seq only (yellow set), RNA-seq only (blue set) and by both methods (gray set) at the whole genome level (left) and expressed exon level (right) in two independent populations A (n = 15, layers) and B (n = 8, broilers). **(C)** Left: Percentages of SNPs in SNP clusters (i.e., 3 or more SNPs in a sliding window of 35 bp, as per GATK definition), in junctions, homopolymers, in population A by DNA-seq only, RNA-seq only and both methods (common). Middle: Number of SNPs detected in 5′ and 3′UTR by gene (*Y*-axis) and the gene number (vertical numbers) in population A. The ratio "RNA-seq specific SNPs/DNA-seq specific SNPs" is indicated at the top of each plot. Right: read supporting SNP distribution at the population level in DNA-seq or RNA-seq data. **(D)** Evolution of the number of detected SNPs as a function of the number of expressed genes using one tissue alone or groups of tissues. Tissues used were liver (figured as a liver), blood (figured as a blood drop) and hypothalamus (figured as a brain).

suggested by GATK (see section "Materials and Methods" and comments below), we found 1,369,740 SNPs. As expected, the number of SNPs detected with RNA-seq is much lower than that in DNA-seq, because only variants present in transcribed regions were detected. Note that the impact of all these filters on the SNP number was provided in the Additional File 3 for DNA-seq and RNA-seq and was quite low, more than 98% of SNP were kept after filtering whatever the population.

To provide a meaningful comparison of both methods, we used the SNPs detected in expressed exons, assessed using RNA-seq with the metric described in section "Materials and Methods." We detected in population A 147,474 expressed exons among the 162,145 exons of the 16,814 expressed genes (on average 8.8 exons per gene). As shown in **Figure 3B** right, in these exons, 85.2% of the 234,500 SNPs detected by DNA-seq were also detected by RNA-seq. In population B, which was composed of only 8 broiler chickens, we found that 65.7% of the SNPs detected with DNA-seq in the expressed exons were also detected by RNA-seq. Assuming SNPs detected by DNA-seq represent the "truth," these percentages represent the sensitivity, or recall, of RNA-seq for SNP detection. This difference in RNA-seq sensitivity between populations A and B is likely due to the number of samples per population (15 versus 8), that affects the extent to which reads at each position are accumulated across the samples (see **Figure 1**).

Concerning the precision of RNA-seq, among the 220,503 SNPs detected by RNA-seq in population A, and the 262,599 SNPs from population B, 90.6 and 91.3%, respectively, were detected by DNA-seq 20X showing a reasonable precision of RNA-seq for the SNP detection. These results are consistent with the findings of Guo et al. (2017), who compared the percentage of SNPs detected using RNA-seq versus exome sequencing and found around 85% concordance. Regarding the 9.4% (20,818 SNPs) RNA-seq specific SNPs, we analyzed different factors that could underlie their detection to highlight those that should be treated with caution (**Figure 3C**) and verify these factors in DNA-seq variants set or in the set of variants called by both methods. We consider the SNPs detected by DNA-seq as true since DNA-seq are now routinely used for SNP detection with the well-proven GATK filters. First, we observed that a large proportion of RNA-seq specific SNPs (46.6%) and DNA-seq specific SNP (40.0%) belonged to a "SNP cluster" (i.e., 3 or more SNPs in a sliding window of 35 bp, as per GATK definition) (**Figure 3C**). This filter is one of the three filters proposed by GATK for RNA-seq SNP detection, but not for DNA-seq detection and the GATK team notes that these filters are not definitive and should be validated by users. Therefore, in the light of these observations, we decided not to remove the "SNP clusters" from our RNA-seq dataset as for DNA-seq dataset, but only to flag them as belonging to a so-called SNP cluster. Indeed, this filter removed 39,783 true SNPs (i.e., True positives detected by both DNA-seq and RNA-seq methods) and consequently the benefit of the precision increase (from 90.6 to 93.5) by removing "SNP clusters" was too small relatively to the recall decrease (from 0.85 to 0.68). The 20,818 RNA-seq specific SNPs can be explained by other factors of lowest impact: (*i*) 5.09% were located at 5 bp or less of an exon-exon junction, *versus* 3.55% for those detected only by DNA-seq; the corresponding ratio, that is significantly greater than 1 (1.4, $p \leq 10^{-17}$, χ^2 test), was expected since RNA-seq deals with spliced transcripts (**Figure 3C**) and therefore RNA-seq read mapping by the aligner is more complicated and more error-prone than DNA-seq read mapping. Since most of them are also observed in DNA-seq, we consider that the SNPs in the vicinity (i.e., 5 bp) of the junctions can be kept, but should be validated by another technique. Note that these SNPs represent only 0.48% of the total SNPs detected by RNA-seq. (*ii*) 3.1% were located in low complexity regions, defined as repetition of at least 5 identical nucleotides, versus 3.4% for the ones detected only by DNA-seq (**Figure 3C**). (*iii*) 2.7 and 5.5 SNPs per gene for RNA-specific SNPs were observed in 5′UTR and 3′UTR regions, respectively, with a fewer 3′UTR SNPs compared to those detected by DNA-seq only (0.5, $p \leq 10^{-16}$, χ^2 test) (**Figure 3C**). This may be due to the fact that mature transcripts undergo exonucleases action, degrading their 3′ extremities and causing their absence in RNA-seq libraries (Gallego Romero et al., 2014). (*iv*) Last, another factor that could be responsible for these RNA-seq specific SNPs is RNA editing, however, according to the literature, it is unlikely that most of the remaining SNPs are due to this mechanism. In mammals, in which RNA editing is well studied, Adenosine-to-Inosine (A-to-I) editing due to ADAR1 and ADAR2 enzymes is the most common editing form and mostly occur in inverted pairs of Alu interspersed repeats (Porath et al., 2014). In chicken *Alu*-like family of interspersed repeats also exist and they are called CR1 (Olofsson and Bernardi, 1983). These editing events tend to occur in clusters, a phenomenon called hyper-editing that introduces $\geq$ 20 mismatches in the sequencing reads (Carmi et al., 2011), that are therefore discarded by the aligner either because of a multi-mapping or no mapping. The prevalence of editing is still discussed: RNA editing is rarely detected when standard mapping filters are used, as shown in mice (Lagarrigue et al., 2013a) and chickens (Frésard et al., 2015; Roux et al., 2016; Shafiei et al., 2019), with less than 200 events, and in humans (Kleinman et al., 2012; Tan et al., 2017) with less than 1000 events per tissue. By contrast, RNA editing is frequently detected when working in repeated regions and rescuing unaligned reads (Picardi et al., 2017). Finally, we observed that SNPs detected only by one method were supported by significantly less reads (either of RNA- or DNA-seq) than the SNPs detected by both methods (**Figure 3C**).

SNP Detection by RNA-Seq: Impact of the Number of Tissues That Are Analyzed

Using blood and hypothalamus samples collected on the same 15 animals (population A), we studied the effect of detecting the SNPs in more than one tissue. RNA-seq from each tissue was not generated at the same time and have been analyzed separately at different occasions. Results are displayed in **Figure 3D**. We detected 1,369,740 SNPs in the liver (as previously stated), 1,481,627 in the blood and 1,511,909 in the hypothalamus, while 16,814 genes were expressed in liver, 16,346 in blood, and 19,733 in hypothalamus. As expected, using combinations of two or three tissues, the number of detected SNPs increased in relation with the number of expressed genes (spearman correlation = 0.96, $p = 3 \times 10^{-3}$) by cumulating the information on all tissues

in which a same gene is more or less expressed. Note that here, we have used our pipeline in a sub-optimal manner, by analysing RNA-seq data per tissue instead of combining the tissues together to increase detection power and reliability. For projects in which RNA-seq from different tissues per animal are all available before SNP detection analysis, we advise users to pool for each animal the RNA-seq files. For SNPs detected in more than one tissue, the concordance between genotypes detected in different tissues was very high, 98.9% without read filtering. Considering genotypes supported by at least 5 reads (respectively 10 reads) the concordance raised to 99.5% (respectively 99.9%).

Genotype (GT) Calling by RNA-Seq Importance of Genotype Call Rate (CR) and Read Depth at the Individual Scale for Selecting SNPs With Enough Reliable Genotypes for *in fine* Calculating Genotype and Allele Frequencies

While reliable SNPs can be detected in the population thanks to some individuals that bear them, it does not necessarily mean that there are enough reads for each individual to produce a genotype (GT). This was exemplified in **Figure 1B** by the brown cells (SNPs 3 and 4), for individuals 4 and 5 ("stress" group) for SNP 3 or most of the individuals of the population for SNP 4. These cases are quite frequent in practice because of gene expression variability between individuals in a given tissue, especially when different conditions are analyzed or also when a SNP is located in an intron of an immature transcript (weakly abundant compared to the mature transcript). Therefore, genotype call rate (CR), defined as the percentage of individuals with a genotype in the population, can be highly variable (e.g., from 16 to 100% in **Figure 1B**, right) from one SNP to another, depending on the number of reads observed in each individual (DP per individual). With 20X DNA-seq data, most of the SNP have a genotype CR close to 100%, as depicted in **Figure 1A**.

These observations indicate that a genotype can be observed with a certain call-rate but its reliability will depend on the DP supporting it. The GT reliability was estimated by the genotype concordance between RNA-seq and DNA-seq, assuming that GT detected by DNA-seq represents the truth. This concordance corresponds to the precision of RNA-seq for GT calling. We tested the RNA-seq precision according to different criteria. First, we conjointly studied in **Figure 4A** the effects of the criteria "genotype CR" and "DP supporting the genotype" on the RNA-seq precision (genotype concordance between RNA-seq and DNA-seq). We found a concordance (of roughly 90%) when no threshold was applied on the DP (purple line); it increased to around 95% for a CR $\geq$ 20% with a DP $\geq$ 5 reads and over 97% for a CR $\geq$ 20% with a DP $\geq$ 10 reads. We then evaluated the impact of the CR alone (without a DP threshold, *x*-axis) versus the CR with a DP $\geq$ 5 reads (*y*-axis), on the genotype concordance between RNA-seq and DNA-seq (solid green isoclines) and on the number of SNPs selected according to the different criteria (dashed blue isoclines) (**Figure 4B**).

Interestingly, only the CR with DP $\geq$ 5 reads have an effect on the genotype concordance and the percentage of selected SNPs, while no such effect is observed for the no DP filtering CR (*x*-axis) comprised between 0 and 50%, as shown by the horizontal isoclines. Hence, we propose for our subsequent analysis on different RNA-seq datasets to select SNPs within the red surface of **Figure 4B** with a (5.reads.DP) genotype CR $\geq$ 20% ensuring a concordance (precision) of almost 95% and a CR $\geq$ 50% ensuring a sufficient number of GT per SNP to calculate the allelic frequencies. We can note that most of the SNPs on this surface have a genotype concordance of more than 97%. We can also note in most of the populations analyzed in the next section that more than 98% of SNPs with (5.reads.DP) genotype CR $\geq$ 20% have a CR $\geq$ 50% (Additional File 4).

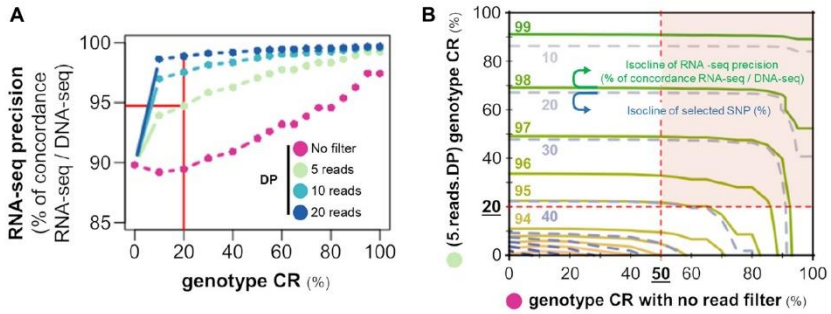


FIGURE 4 | The passage from SNP to GT necessitates a read depth threshold. **(A)** Evolution of the percentage of genotype concordance between RNA-seq and DNA-seq (*y*-axis) for the 15 RpRm birds as a function of genotype call rate in the population (CR: *x*-axis) supported by at least 5 (light green), 10 (light blue), or 20 (dark blue) reads or without read filter (purple curve). The red lines correspond to the criteria used in the further analysis (CR $\geq$ 20% with a DP $\geq$ 5 reads) and the corresponding RNA-seq precision. **(B)** Isoclines of the percentage of genotype concordance between RNA-seq and DNA-seq (solid green lines) and of the percentage of SNPs selected out to the original set (dashed blue lines) according to the CR with no read filter (*x*-axis) and the CR with at least 5 reads [(5.reads.DP) genotype CR (%), *y*-axis]. Red surface: SNPs selected after filtering on (5.reads.DP) genotype CR $\geq$ 20% and a CR $\geq$ 50%.

Number of SNPs and Genotypes Detected by RNA-Seq in 11 Populations

As shown in **Table 1** which gives an overview of the SNP diversity in 11 chicken populations, we detected between 1.1 and 3.8 M SNPs per population using liver RNA-seq datasets. Using all the tissues available (1–5 tissues depending on the population), we detected more SNPs, consistently with our previous result (see **Figure 3D**): between 1.7 and 5.5 M SNPs with a fold increase of × 1.18 to × 2.48 depending on the number and nature of analyzed tissues. Across populations and using all tissues, we found a total of 9.5 M SNPs having at least one alternative allele in at least one population (SNP union), and 241,960 SNPs that had at least one alternative allele in each of the 11 populations (SNP intersection). The union of our SNPs contains 23% (2,175,528) yet-unreported SNPs in the reference Ensembl v94 dbSNP database [(Ensembl, 2018): 21 M SNPs]. The intersection of our SNPs contains 5.1% (12,203 SNPs) of the SNPs present in the 600K genotyping array (Kranis et al., 2013).

We then filtered SNPs on genotype call rate and read depth (**Table 1**, "Selected GT") and found between around 0.4 and 1.7 M SNPs using all tissues, 37% of the SNPs observed previously. These results on 11 populations show that a large number of SNPs (two thirds) were detected at the population level thanks to the accumulation of reads across all individuals of the population, but that within each individual, read counts are not sufficient to reliably determine a genotype. Nevertheless, the number of SNPs with a genotype per population remains in the order of magnitude of several hundred thousand to a few millions with a union of 3.3 M and an intersection of 73,223 SNPs. In the liver, for which data was available in all but one population (Rmx6), the union and intersection are of the same order of magnitude: 1.7 M and 67,341 SNPs, respectively. After selecting for a MAF (minor allele frequency) $\geq$ 10% in order to discard rare SNPs or those resulting from sequencing errors, the number of SNPs was halved in all populations with a grand total of 2.2 and 1.3 M for the multi-tissue and liver union, respectively. As expected, the intersection drastically decreased to approximatively 2,000 SNPs, since this set corresponds to the SNPs with a MAF $\geq$ 10% in each of the 11 populations. The list of the 9.5 M of SNPs including 3.3 M with a GT and 2.2 M with MAF $\geq$ 10% is available on http://www.fragencode.org/lnchickenatlas.html.

Rare Deleterious Variants Detection in the Populations

We predicted the impacts of the 9,496,283 SNPs detected in at least one population using the VEP tool (McLaren et al., 2016) which predicts the potential consequences of the SNPs in each of the transcripts carrying them: we found 33,304,412 consequences. As expected, the vast majority of the SNPs affected non-coding regions (**Figure 5A**) and among the 472,319 SNPs affecting a coding-region, a majority were synonymous (63%) or non-deleterious missense (28%) as shown in **Figure 5B**.

Among all these predictions, we focused on the predicted consequences with the most severe putative impacts as defined by the gnomAD consortium, which only considers the PCG (Protein Coding Genes) (Genome Aggregation Database Consortium et al., 2020): variants in the splice regions, start and stop codon loss or stop codon gain even if the severity of the latter depends on its position in the coding sequence. We also added missense variants with a SIFT score $\leq$ 0.05. As reported by gnomAD (The GTEx Consortium, 2020), these SIFT-deleterious SNPs generally have a low frequency in the populations and can be mistaken for sequencing errors. Hence, it is crucial to select SNPs with genotypes (as defined previously) and a MAF $\geq$ 10% in at least one population (i.e., the ALT allele observed for example at least 4 times in a population of 16 individuals as for FAyo and LSnu populations) to make sure that the deleterious allele is not spurious. Thanks to our data from 382 individuals from the 11 populations, we listed a total of 25,344 strong predicted impacts (**Figure 5C**), corresponding to 14,496 SNPs and 67,58 genes, among them were 590 predictions of stop gained (404 genes), 8,126 of a coding or non-coding gene splice site change (donor and acceptor), 16,307 SIFT-predicted deleterious missenses and 321 other predictions (start lost, stop lost). Out of these 25,344 deleterious-predicted impacts, we found 5,654 (22%) predictions corresponding to 2,872 (20% of 14,496 SNPs) variants in 1,884 genes for which the homozygous ALT/ALT genotype was absent, in all populations in which the ALT allele was detected and, respectively, 7,740 (31%) predictions corresponding to 4,072 (28% of 14,496 SNPs) variants in 2,515 genes with ALT/ALT frequency $\leq$ 5%. The analysis of tolerated missense SNP show that the higher the SIFT score (i.e., tolerated variant), the lower the percentage of SNP with a low frequency ($\leq$ 5%) of ALT/ALT genotype (**Figure 5D**). The same analysis performed with 217,119 synonymous variants showed lower percentages with 9% SNPs with ALT/ALT genotype absent and 13% SNP with ALT/ALT frequency $\leq$ 5%. Such results are compatible with a homozygous state which is lethal or strongly negatively selected (28 versus 13%, $p \leq 10^{-20}$, χ^2 test), suggesting an important role for the genes associated to these variants with severe-predicted impact. Such variants obtained using RNA-seq data constitute a new complementary resource to Ensembl dbSNP allowing to explore variants (deleterious or not) according to their genotypic and allelic frequencies in different populations of a farm species. For example, two deleterious missense SNPs (SIFT-score = 0) are presented in **Figure 6**. One is already reported in dbSNP (Ensembl genome browser 94, 2020) and affects XBP1 protein by changing a positive charged amino acid (Arginine, R) into an aromatic and hydrophobic amino acid (Tryptophan, W) (**Figure 6A**). This SNP is observed in two of the ten analyzed populations, FLLL and Novo2, with 5 and 10 heterozygous birds among 48 and 40 animals analyzed, respectively, whereas no ALT/ALT homozygous birds were observed (**Figure 6B**). This gene is ubiquitously expressed in chicken as in human (**Figure 6C**). It codes the "Tax-Responsive Element-Binding Protein 5" transcription factor which has important cellular and physiological roles related to the "unfolded protein response" pathway in the endoplasm reticulum [(Lee et al., 2003) and for review (Glimcher et al., 2020)] and also to hepatic insulin resistance (Zhou et al., 2011).

The second SNP, not reported in dbSNP, affects the SERGEF protein (alias DelGEF) by changing an aromatic, hydrophobic and positive charged amino acid (Histidine, H) into an

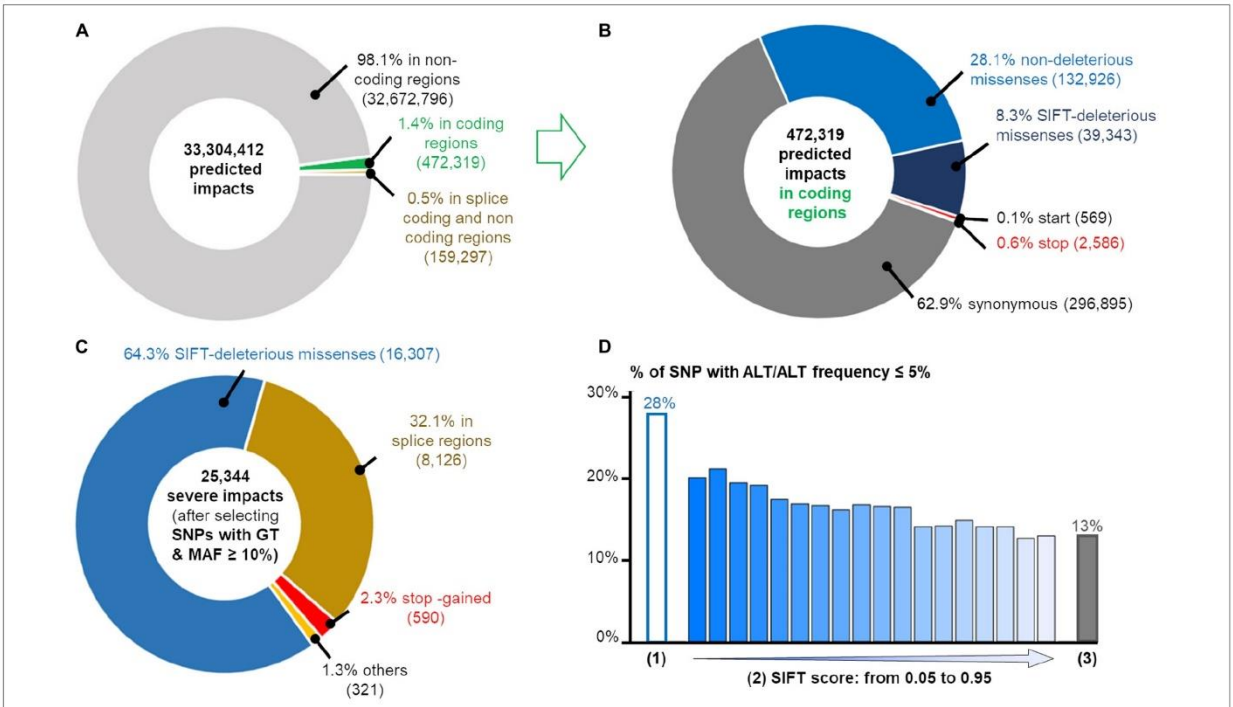


FIGURE 5 | Annotation of 9,496,283 SNPs using Variant effect Predictor (VEP) (McLaren et al., 2016). **(A)** Distribution of variant effect predictions among non-coding (light gray), splice regions related to coding and non-coding genes (orange) and coding (green) regions. **(B)** SNP annotation in coding regions: synonymous (dark gray), non-deleterious (light blue) and deleterious (blue) missenses, and consequences affecting stop (red) and start (orange) codons. Total number of consequences are indicated between parentheses. **(C)** Annotation of SNPs predicted as deleterious and filtered according to the GT criteria (as defined previously) and a MAF ≥ 10% in at least one population. **(D)** Percentage of SNP with a very low frequency (≤ 5%) of ALT/ALT genotype for three SNP sets: (1) = the 25,344 deleterious SNP described on the left; (2) = the tolerated SIFT-missense SNP according to the SIFT score and (3) = the synonymous SNP set. The splice sites correspond to the donor or acceptor splice sites of coding and long non-coding genes.

unchanged amino acid (Tyrosine, Y) (**Figure 6A**). This SNPs was observed in two populations, LSnu and Fayoumi, with 6 and 3 heterozygous birds among 16 animals, respectively, whereas no ALT/ALT homozygous birds were observed. This gene is also relatively ubiquitously expressed in chicken as in human (**Figure 6C**). The functions of this gene, which codes the "Secretion Regulating Guanine Nucleotide Exchange Factor" seem to be poorly known: 9 publications found in PubMed with the key words, SERGEF or DELGEF. As illustrated by these two examples (XBP1 and SERGEF), the analysis of various populations allowed to increase the number of rare deleterious variants detected.

Potential for Allele-Specific Expression Analysis in Various Populations

Allele-specific expression (ASE) analysis requires a heterozygous SNP in the expressed feature, to test an eventual imbalance in the expression between the two parental chromosomes. Usually, the expression is evaluated using RNA-seq and the SNPs are detected using DNA-seq, which is expensive when working on a dozen or more individuals. Since we have shown that RNA-seq allows detecting a large number of reliable SNPs in expressed regions, we studied in this section, the potential of RNA-seq data for performing ASE analysis. To this end, the **Figure 7A** provides the average numbers of genes across various populations, having at least one SNP with different filters (SNPs with an associated GT, a MAF ≥ 10% and an heterozygous status in at least 25% of the population). We also indicated the average SNP number per gene (column "S/g") to give an idea of the RNA-seq potential to test ASE along the gene. We indicated the results for two types of genes: the protein-coding genes (PCG) and the long non-coding genes (lncRNA), which are increasingly considered as important regulators of gene expression but are also known to be less expressed than PCG (Derrien et al., 2012; Muret et al., 2017; Le Béguec et al., 2018). This is the reason why we studied two expression thresholds: 0.1 and 1 TPM commonly used when working on lncRNA and PCG, respectively. Finally, results in **Figure 7A** are presented either for SNPs detected in exons (i.e., mature transcripts) (top) or for SNPs detected in exons or introns hence including immature transcripts (bottom).

The first key result is that the number of genes with at least one SNP are similar in both cases (exons only versus exons + introns), meaning that there are enough SNPs to study ASE in exonic regions only, i.e., mature transcript, despite a much lower number of SNPs per gene when SNPs are

A

gnName	chr	pos	ref	alt	SNPid	SIFTsc.	codon	aa	position
XBP1	15	7763534	G	A	rs738179721	0	Cgg/Tgg	R/W	105/255
SERGEF	5	12450547	G	A	NA	0	Cat/Tat	H/Y	217/298

B

gnName	pop.	# ind.	allele frequencies		genotype frequencies		
			ref	alt	ref/ref	ref/alt	alt/alt
XBP1	FLLL	48	0.95	0.05	0.9	0.1	0
XBP1	Novo2	40	0.875	0.125	0.75	0.25	0
SERGEF	LSnu	16	0.815	0.185	0.63	0.37	0
SERGEF	FAyo	16	0.905	0.095	0.81	0.19	0

C

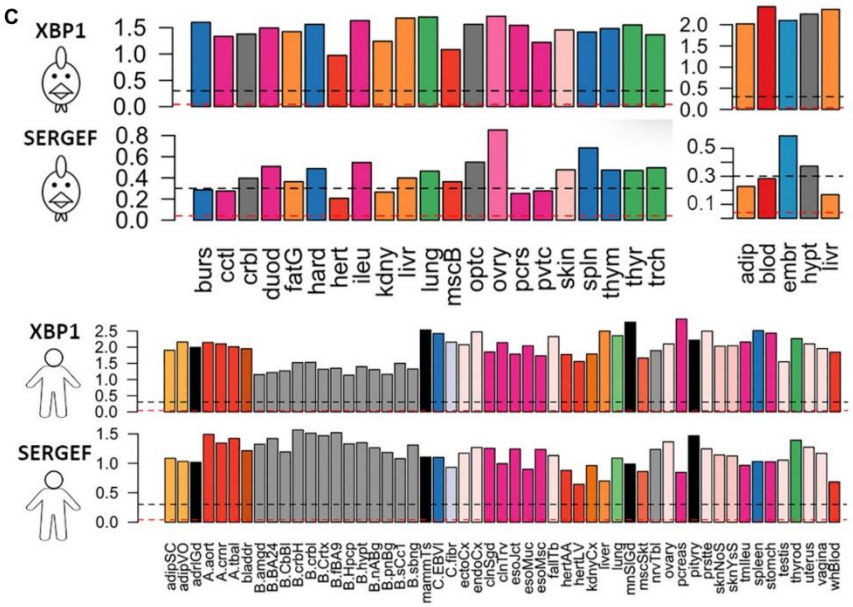


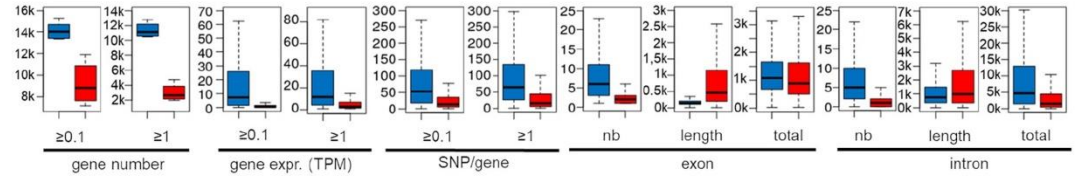
FIGURE 6 | Two examples of deleterious missense SNPs impacting two protein coding genes (XBP1 and SERGEF). **(A)** genomic position of the SNP with its identifier (SNPid) in Ensembl dbSNP and its impact on the protein with SIFTsc.: SIFT score, codon/modified codon, amino acid/modified amino acid and its position in the protein. **(B)** pop.: population with the individual size (# ind.) observed per population and the frequencies of the alleles and genotypes. **(C)** Tissue expressions [log10(TPM + 1)] in chicken using two datasets composed of 21 tissues (ERP014416) (left) and 5 tissues RpRm population) (right) and in human through the 53 tissues from the GTEX consortium (The GTEx Consortium, 2020). Abbreviations for the 21-tissue dataset: burs, bursa of Fabricius; cctl, cecal tonsils; crbl, cerebellum; duod, duodenum; fatG adipose tissue around the gizzard; hard, harderial gland; hert, heart; ileu, ileum; kdny, kidney; livr, liver; lung, lung; mscB breast muscle; optc, optical lobe; ovry, ovary; pcrs, pancreas; pvtc, proventriculus; skin, skin; spln, spleen; thym, thymus; thyr, thyroid gland; trch, trachea; and for the 5-tissue dataset: adip, abdominal adipose tissue; blod, blood; embr, 4.5 day embryos; hypt, hypothalamus; livr, liver; for more details in these 3 datasets and associated samples see Jehl et al. (2020). Black dashed line: gene expression with TPM $\geq$ 1 and red dashed line: TPM $\geq$ 0.1.

only selected in exons (**Figures 7A,B**). When working with exonic SNPs, there are on average 17–28 SNPs without filter (8–10 SNPs after all filters) per gene showing the possibility to test ASE along genes. Despite a lower exonic length in lncRNA compared to the PCG (**Figure 7B**), this number is higher for lncRNA compared to PCG (22–28 versus 15–17) probably due to lower selective pressure on lncRNA compared to PCG. The second key result, after applying 2 filters (GT and MAF $\geq$ 10%), is that 81% of PCG (9,232) and 68% of lncRNA (2,028) expressed at TPM $\geq$ 1 are analyzable for ASE. These numbers decreased a little after applying an additional filter related to the heterozygosity percentage, with 72% of PCG and 56% of lncRNA (i.e., about 10,000 genes). The variability of this "ASE analyzable genes" percentage is moderate (Additional File 5): on average 72% from 65 to 89% with an except for the "RpRm" (48%) probably due to its high consanguinity and its large size, the filter of 25% of heterozygosity impacting more the populations with a larger sample size. The same tendencies regarding the percentage of genes that can be analyzed were observed for the PCG (TPM $\geq$ 0.1) and for lncRNA (both for TPM $\geq$ 0.1 and $\geq$ 1) (Additional File 5). We can note that the selected lncRNA percentage satisfying the filters is always lower than the selected PCG percentage ($-$15% for genes with an expression $\geq$ 1TPM and $-$30% for genes with an expression $\geq$ 0.1TPM). This is mainly due to the lower expression of lncRNA compared to PCG (Jehl et al., 2020; **Figure 7B**), despite higher sequence variability for the former.

A

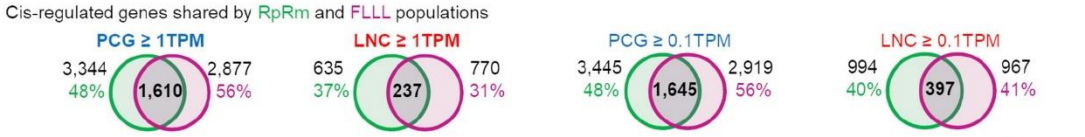
expr. threshold	biotype	expr. gene	no filter SNP	no filter gene	no filter S / g	no filter %	GT filter SNP	GT filter gene	GT filter S / g	GT filter %	MAF ≥ 10% SNP	MAF ≥ 10% gene	MAF ≥ 10% S / g	MAF ≥ 10% %	Het ≥ 25% SNP	Het ≥ 25% gene	Het ≥ 25% S / g	Het ≥ 25% %
SNP in exons																		
1 TPM	PCG	11 384	181 513	10 770	17	95	144 895	10 522	14	92	89 263	9 232	10	81	67 591	8 281	8	72
	LNC	2 982	72 040	2 608	28	88	46 401	2 365	20	79	27 727	2 028	14	68	17 756	1 701	10	56
0.1 TPM	PCG	14 183	212 598	13 116	16	92	153 716	11 560	13	82	94 251	10 063	9	71	70 063	8 792	8	62
	LNC	9 228	175 253	7 899	22	86	70 903	4 517	15	49	42 108	3 753	11	40	25 616	2 893	8	31
SNP in genes (i.e. exons + introns)																		
1 TPM	PCG	11 384	1 087 288	11 098	97,6	98	390 248	10 750	36	94	242 720	9 725	25	85	156 265	8 701	17	76
	LNC	2 982	110 267	2 670	41,7	90	59 006	2 403	24	81	35 356	2 080	17	69	22 408	1 756	12	58
0.1 TPM	PCG	14 183	1 166 864	13 541	85,6	96	403 222	11 825	34	84	250 173	10 586	23	75	159 993	9 246	17	65
	LNC	9 228	252 298	8 060	31,1	87	91 564	4 612	19	50	54 772	3 869	14	42	33 193	3 006	11	32

B



C Percentage of cis-regulated genes among the expressed genes in liver

expr. threshold	biotype	pop: RpRm expr.gene	Nb of individuals with ASE genes 2	3	4	pop: FLLL expr.gene	Nb of individuals with ASE genes 2	3	4	Mean 2
1 TPM	PCG	9748	34.3	26.2	21.0	11707	24.6	18.0	13.7	29.5
	LNC	1828	34.7	27.8	23.4	3528	21.8	15.7	11.7	28.3
0.1 TPM	PCG	11254	30.6	23.3	18.7	12178	24.0	17.6	13.3	27.3
	LNC	4265	23.3	18.1	15.1	5326	18.2	13.0	11.7	20.8



D

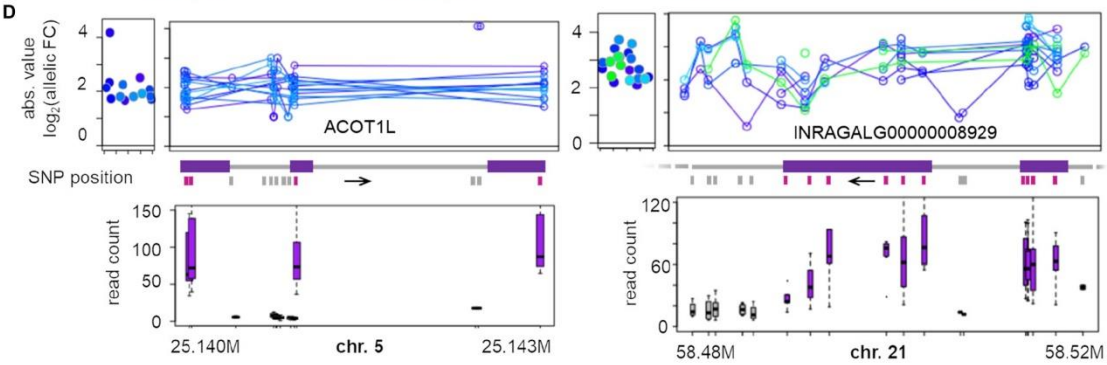


FIGURE 7 | Overview of the analyzable genes for allele-specific expression in the liver of various populations for two gene biotypes, Protein Coding Gene (PCG) and Long Non-Coding gene (lncRNA), at two gene expression thresholds (0.1 TPM and 1 TPM) and for 3 filters. **(A)** Average numbers for all populations analyzed here. These average numbers are provided for both PCG (blue) and lncRNA (red) biotype, with minimum expression of 0.1 or 1 TPM ("expr. threshold"), and considering only the SNPs in exons (top part) or in the whole gene, i.e., in both exons and introns (bottom part). **(B)** Feature of lncRNA and PCG. **(C)** Percentage of gene with a significant allele specific expression in two populations RpRm (in left) and FLLL (in right) in comparison to the expressed gene number. Venn diagrams provide the number of ASE genes (in at least 2 individuals) shared by RpRm and FLLL populations. **(D)** Overview of the ASE of ACOT1L (left) and INRAGALG00000008929 (right). For each ASE sample, absolute values of the log2 allelic fold-change are represented at the gene-level (left of the panels) and for each SNP located in the haplotype used by phASER (right). Boxplot of the read number associated to each SNP are represented (bottom), in purple for the SNP located in exons and in gray for those in introns. FC, fold-change; chr, chromosome.

Cis-Regulated Genes in the Liver of Two Populations

To provide an estimation of the number of *cis*-regulated genes in one tissue, we performed an ASE analysis of the liver samples of the RpRm and the FLLL populations using phASER and its downstream tool, phASER Gene AE, that phase SNPs at the gene level (see also section "Materials and Methods"). Using exonic and intronic SNPs and selecting genes having one haplotype with at least 10 reads, we found for genes with an hepatic expression ≥ 1 TPM, that in average 29%

of the expressed PCG or lncRNA genes were *cis*-regulated (∼34% for RpRm and ∼23% for FLLL) (**Figure 7C**). For lncRNA with hepatic expression ≥ 0.1 TPM which represents most of this biotype, we found a lower percentage of *cis*-regulated genes (21%) because they are less expressed and some of them did not have more than 10 reads for at least one "super-allele" analyzed by phASER (see section "Materials and Methods"). Interestingly, among these *cis*-regulated genes, ∼50% and 37% are shared by both populations for the protein-coding genes and long non-coding genes, respectively (**Figure 7C**). Two examples of *cis*-regulated genes are provided in **Figure 7D** with a PCG, ACOT1L (ENSGALG00000008752), and a lncRNA, INRAGALG000000089295. Overall, these numbers are consistent with the literature: Zhuo et al. (2017) found that 15% of the genes were *cis*-regulated in chicken embryo liver, and Lagarrigue et al. (2013b) found a similar number in mice liver. In humans, the GTEx consortium (The GTEx Consortium, 2020) found that 26% (4,415) of the expressed genes (17,243) were *cis*-regulated in the liver.

Diversity Exploration Using RNA-Seq Variants

Finally, we explored genetic links between populations using the genotypic frequencies of SNPs detected by RNA-seq, which represent a set of SNPs, which may be under a larger selective pressure than those used in genotyping SNP chips. Indeed, the latter are considered as having a neutral effect, while most the SNPs present in our data are located in expressed regions and affect proteins to some extent (from almost neutral synonymous to deleterious stop gained).

The classification in **Figure 8** was produced using the intersection of SNPs with GT of the 10 populations with a liver presented in **Table 1** (67,341 SNP set). This classification is consistent with the known chicken population history, indicating that these SNPs detected by RNA-seq and their associated genotypes allow distinguishing different populations. The classification separated clearly the RJFh (red circle arc with a Red Jungle Fowl population, used here to represent the "ancestral" population), then the broilers (blue circle arc), the brown-egg layers (dark green circle arc), and the cream- or white-egg layers (brown circle arc with Fayoumi breed and Fr-Ag population which is an experimental leghorn line). We also observed the expected sub-groups within these 3 types of populations: the commercial lines (Novo1 and Novo2 for the layers, Cobb and HerX for the broilers) separated from the experimental lines (RpRm for the brown-egg layers, FLLL for the broilers). Interestingly for these 2 last populations, this SNP set shows a clear distinction between two subpopulations that have been divergently selected for a specific trait: Rp and Rm divergent for the residual feed intake and FL and LL divergent for body fat whereas the two Novogen populations (Novo1 and Novo2) are not distinct. We can note that the SNPs predicted as "missense" by VEP and "deleterious" by SIFT provide the same classification between the populations as the one shown in **Figure 8** (data not shown).

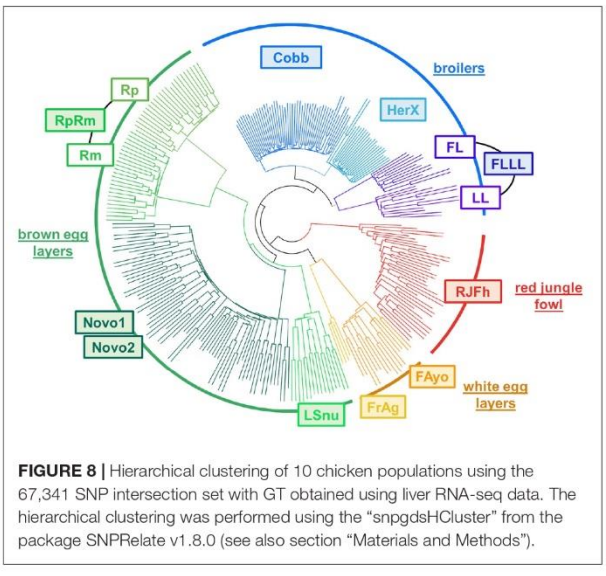


FIGURE 8 | Hierarchical clustering of 10 chicken populations using the 67,341 SNP intersection set with GT obtained using liver RNA-seq data. The hierarchical clustering was performed using the "snpgdsHCluster" from the package SNPRelate v1.8.0 (see also section "Materials and Methods").

CONCLUSION

We show here that RNA-seq data, which are cheaper to generate and store compared to DNA-seq data, can be a reliable resource for performing different analyses based on polymorphism detection. By comparing DNA-seq and RNA–seq results generated from the same animals in two independent chicken populations, this study provides a workflow to produce reliable SNPs and genotypes from RNA–seq data. We ran through this pipeline 767 RNA–seq of 382 birds from 11 populations and provided a per-population estimation of the average genotyped SNPs count per tissue (more than 550,000) and an overview of the predicted consequences of SNPs located in coding regions. In particular, thanks to this large RNA-seq dataset, we identified 440 genes containing a stop-gained impact, known to be rare because of their potentially severe impact, especially when located in the first third of the coding sequences (133 genes). In a companion study (Degalez et al., submitted), we checked the possible existence of more than one SNP in a given codon, that could "rescue" a stop-gained situation. We then gave an overview across 11 populations of genes that could be analyzed for ASE, i.e., having at least one SNP allowing to distinguish expression from both chromosomes. We applied phASER on liver RNA-seq data of two populations and identified around 21 to 30% of *cis*-regulated genes depending on the analyzed population and the gene biotype (PCG versus lncRNA), these results were consistent with other studies conducted in other species.

This study represents a first step to more ambitious projects that could analyze tens of thousands of available RNA-seq datasets to build a GTEx-like atlas reporting *cis*- and *trans*- genetic associations with gene expression, as previously performed in human (The GTEx Consortium, 2020) and more recently in cattle (Liu et al., 2020).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because all data have been obtained from public databases as described in the materials and methods, section "DNA-Seq and RNA-Seq Data."

AUTHOR CONTRIBUTIONS

CK and SLa conceived the study and coordinated the study. FJ, LL, CD, BA, MT-B, BB, TB, DG, HA, SF, SD, EG, FP, TZ, and SLa participated to the set-up of the experimental design and sample collection. LL, CD, and SLe carried out all RNA extractions. OB generated RNA-seq libraries and sequencing. FJ, FD, MB, CK, and SLa performed bioinformatic processing of the RNA-seq data and carried out the whole analyses. FL, PB, and MC participated to bioinformatic analyses. FJ, FD, and SLa drafted the manuscript. MB, FL, BA, MT-B, BB, HA, SF, SD, EG, TZ, FP, and CK helped to improve the manuscript. All authors read and approved the final version.

FUNDING

The sample and/or data were collected in the frame of projects that received financial support from the European Union's H2020 Program under Grant Agreement No. 633531 (Feed-a-Gene project), the French National Agency of Research (FatInteger project ANR-11-SVS7; ChickStress project, ANR-13-ADAP; EpiBird ANR project PCS-09-GENM-010), from French institutions as Institut Agro-AGROCAMPUS OUEST and INRAE [Fr-AgENCODE project (2015-2017), ELASETIC project (2012)]. This work was also supported by France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (contrat ANR-10-INBS-09). FJ and FD are Ph.D. fellows supported by the Brittany region (France) and the INRAE (Animal Genetics Division). These funding bodies had no role in the design of the study, in the collection, analysis, and interpretation of data, or in writing the manuscript.

ACKNOWLEDGMENTS

We are grateful to the GeT-PlaGe platform (doi: 10.15454/1.5572370921303193E12) for RNA-seq libraries and sequencing and the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi: 10.15454/1.5572369328961167E12) for providing help and/or computing and/or storage resources. A warm thanks to Morgane Boutin who passed away this year and who did a lot in RNA extraction.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.655707/full#supplementary-material

Additional File 1 | Detail per population of the number of birds, tissues and samples.

Additional File 2 | Boxplot of the exon expression in log10 (RpKb+1) for both RpRm and FLLL populations compared to noise. Noise corresponds to the expression of a set of randomly selected loci in the genome.

Additional File 3 | Effect of GATK filters on SNP numbers with RNA-seq and DNA-seq data. For RNA-seq, the analyses were performed in different tissues and populations. For DNA-seq, the analyses were performed for the two populations, RpRm with 15 individuals and FLLL with 8 individuals. Abbreviations for the tissue: adip: abdominal adipose tissue, blod: blood, embr: embryos, hypt: hypothalamus, livr: liver, spln: spleen. "rawSNP": SNP detected before the filters' application. "filtered biallelic SNP": SNP detected after the filters' application.

Additional File 4 | SNP counts per population with each genotype filters independently.

Additional File 5 | Overview of the analyzable genes for allele-specific expression in the liver of various populations for Protein Coding Gene (PCG) and Long Non-Coding gene (lncRNA), at two gene expression thresholds (0.1 TPM and 1 TPM) with the mean across the populations ("mean") and the coefficient variation ("CV"), i.e., standard deviation divided by mean. "expr. gene": expressed genes, "gene": genes bearing the SNP, "S/g": number of SNPs divided by the number of gene, "%": proportion of genes bearing the SNPs relative to the expressed genes. The 4 filters: no filter, GT satisfying our criteria ("GT filter"), with MAF $\geq$ 10% ("MAF $\geq$ 10%") and heterozygosity rate $\geq$ 25% ("Het. $\geq$ 25%").

REFERENCES

Adetunji, M. O., Lamont, S. J., Abasht, B., and Schmidt, C. J. (2019). Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. *PLoS One* 14:e0216838. doi: 10.1371/journal.pone.0216838

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., et al. (2013). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24, 14–24. doi: 10.1101/gr.155192.113

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bordas, A., Tixier-Boichard, M., and Merat, P. (1992). Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. *Br. Poult. Sci.* 33, 741–754. doi: 10.1080/00071669208417515

Carmi, S., Borukhov, I., and Levanon, E. Y. (2011). Identification of widespread ultra-edited human RNAs. *PLoS Genet.* 7:e1002317. doi: 10.1371/journal.pgen.1002317

Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y., and Lappalainen, T. (2016). Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.* 7:12817.

Chamberlain, A. J., Vander Jagt, C. J., Hayes, B. J., Khansefid, M., Marett, L. C., Millen, C. A., et al. (2015). Extensive variation between tissues in allele specific

expression in an outbred mammal. *BMC Genomics* 16:993. doi: 10.1186/s12864-015-2174-0

Cornwell, M., Vangala, M., Taing, L., Herbert, Z., Köster, J., Li, B., et al. (2018). VIPER: visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics* 19:135. doi: 10.1186/s12859-018-2139-9

Deelen, P., Zhernakova, D. V., de Haan, M., van der Sijde, M., Bonder, M. J., Karjalainen, J., et al. (2015). Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* 7:30.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Ensembl (2018). *dbSNP - Gallus Gallus 5 - V94*.

Frésard, L., Leroux, S., Roux, P.-F., Klopp, C., Fabre, S., Esquerré, D., et al. (2015). Genome-wide characterization of RNA editing in chicken embryos reveals common features among vertebrates. *PLoS One* 10:e0126776. doi: 10.1371/journal.pone.0126776

Frésard, L., Leroux, S., Servin, B., Gourichon, D., Dehais, P., Cristobal, M. S., et al. (2014). Transcriptome-wide investigation of genomic imprinting in chicken. *Nucleic Acids Res.* 42, 3768–3782. doi: 10.1093/nar/gkt1390

Fu, W., Dekkers, J. C., Lee, W. R., and Abasht, B. (2015). Linkage disequilibrium in crossbred and pure line chickens. *Genet. Select. Evol.* 47:11. doi: 10.1186/s12711-015-0098-4

Gallego, Romero I, Pai, A. A., Tung, J., and Gilad, Y. (2014). RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* 12:42. doi: 10.1186/1741-7007-12-42

Genome Aggregation Database Consortium Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443.

GitLab (2019). *Snakemake/1000RNASeq_chicken/calling · master · bios4biol / workflows*. Available online at: https://forgemia.inra.fr/bios4biol/workflows/tree/master/Snakemake/1000RNASeq_chicken/calling (accessed August 28, 2019).

Glimcher, L. H., Lee, A.-H., and Iwakoshi, N. N. (2020). XBP-1 and the unfolded protein response (UPR). *Nat. Immunol.* 21, 963–965. doi: 10.1038/s41590-020-0708-3

Gondret, F., Vincent, A., Houée-Bigot, M., Siegel, A., Lagarrigue, S., Causeur, D., et al. (2017). A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. *BMC Genom.* 18:244. doi: 10.1186/s12864-017-3639-0

Guo, Y., Zhao, S., Sheng, Q., Samuels, D. C., and Shyr, Y. (2017). The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. *BMC Genom.* 18:690. doi: 10.1186/s12864-017-4022-x

Jehl, F., Désert, C., Klopp, C., Brenet, M., Rau, A., Leroux, S., et al. (2019). Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. *BMC Genom.* 20:1033. doi: 10.1186/s12864-019-6384-8

Jehl, F., Muret, K., Bernard, M., Boutin, M., Lagoutte, L., Désert, C., et al. (2020). An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci. Rep.* 10:20457.

Kleinman, C. L., Adoue, V., and Majewski, J. (2012). RNA editing of protein sequences: a rare event in human transcriptomes. *RNA* 18, 1586–1596. doi: 10.1261/rna.033233.112

Koster, J., and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi: 10.1093/bioinformatics/bts480

Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., et al. (2013). Development of a high density 600K SNP genotyping array for chicken. *BMC Genom.* 14:59. doi: 10.1186/1471-2164-14-59

Krueger, F. (2021). *FelixKrueger/TrimGalore*. Available online at: https://github.com/FelixKrueger/TrimGalore (accessed March 15, 2021).

Lagarrigue, S., Hormozdiari, F., Martin, L. J., Lecerf, F., Hasin, Y., Rau, C., et al. (2013a). Limited RNA editing in exons of mouse liver and adipose. *Genetics* 193, 1107–1115. doi: 10.1534/genetics.112.149054

Lagarrigue, S., Martin, L., Hormozdiari, F., Roux, P.-F., Pan, C., van Nas, A., et al. (2013b). Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis -eQTL identified using genetic linkage. *Genetics* 195, 1157–1166. doi: 10.1534/genetics.113.153882

Le Béguec, C., Wucher, V., Lagoutte, L., Cadieu, E., Botherel, N., Hédan, B., et al. (2018). Characterisation and functional predictions of canine long non-coding RNAs. *Sci. Rep.* 8:13444.

Lee, A.-H., Iwakoshi, N. N., and Glimcher, L. H. (2003). XBP-1 regulates a subset of endoplasmic reticulum resident chaperone genes in the unfolded protein response. *Mol. Cell Biol.* 23, 7448–7459. doi: 10.1128/mcb.23.21.7448-7459.2003

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291.

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12:323. doi: 10.1186/1471-2105-12-323

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [Preprint]. Available online at: http://arxiv.org/abs/1303.3997 (accessed April 6, 2021).

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., et al. (2020). A comprehensive catalogue of regulatory variants in the cattle transcriptome. *bioRxiv* [Preprint]. doi: 10.1101/2020.12.01.406280v1

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., et al. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–777. doi: 10.1038/nature08903

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Muret, K., Klopp, C., Wucher, V., Esquerré, D., Legeai, F., Lecerf, F., et al. (2017). Long noncoding RNA repertoire in chicken liver and adipose tissue. *Genet. Sel. Evol.* 49:6.

Ng, P. C. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Oikkonen, L., and Lise, S. (2017). Making the most of RNA-seq: pre-processing sequencing data with Opossum for reliable SNP variant detection. *Wellcome Open Res.* 2:6. doi: 10.12688/wellcomeopenres.10501.2

Olofsson, B., and Bernardi, G. (1983). The distribution of CR1, an Alu-like family of interspersed repeats, in the chicken genome. *Biochim. Biophys. Acta Gene Struct. Express.* 740, 339–341. doi: 10.1016/0167-4781(83)90143-4

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Peng, Z., Cheng, Y., Tan, B. C.-M., Kang, L., Tian, Z., Zhu, Y., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* 30, 253–260. doi: 10.1038/nbt.2122

Picardi, E., D'Erchia, A. M., Lo Giudice, C., and Pesole, G. (2017). REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* 45, D750–D757.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772. doi: 10.1038/nature08872

Piskol, R., Ramaswami, G., and Li, J. B. (2013). Reliable identification of genomic variants from RNA-Seq data. *Am. J. Hum. Genet.* 93, 641–651. doi: 10.1016/j.ajhg.2013.08.008

Porath, H. T., Carmi, S., and Levanon, E. Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat. Commun.* 5:4726.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., et al. (2013). Development of strategies for SNP detection in RNA-Seq data: application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS One* 8:e58815. doi: 10.1371/journal.pone.0058815

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Roux, P.-F., Boitard, S., Blum, Y., Parks, B., Montagner, A., Mouisel, E., et al. (2015). Combined QTL and selective sweep mappings with coding SNP annotation and cis-eQTL analysis revealed PARK2 and JAG2 as new candidate genes for adiposity regulation. *G3 Genes Genomes Genet.* 5, 517–529. doi: 10.1534/g3.115.016865

Roux, P.-F., Frésard, L., Boutin, M., Leroux, S., Klopp, C., Djari, A., et al. (2016). The extent of mRNA editing is limited in chicken liver and adipose, but impacted by tissular context, genotype, age, and feeding as exemplified with a conserved edited site in COG3. *G3* 6, 321–335. doi: 10.1534/g3.115.022251

Savary, C., Kim, A., Lespagnol, A., Gandemer, V., Pellier, I., Andrieu, C., et al. (2020). Depicting the genetic architecture of pediatric cancers through an integrative gene network approach. *Sci. Rep.* 10:1224.

Shafiei, H., Bakhtiarizadeh, M. R., and Salehi, A. (2019). Large-scale potential RNA editing profiling in different adult chicken tissues. *Anim. Genet.* 50, 460–474. doi: 10.1111/age.12818

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–132. doi: 10.1038/nrg3642

Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. *Nature* 550, 249–254.

Tang, X., Baheti, S., Shameer, K., Thompson, K. J., Wills, Q., Niu, N., et al. (2014). The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. *Nucleic Acids Res.* 42:e172. doi: 10.1093/nar/gku1005

The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. doi: 10.1126/science.aaz1776

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 43, 11.10.1–11.10.33.

Wang, C., Davila, J. I., Baheti, S., Bhagwate, A. V., Wang, X., Kocher, J.-P. A., et al. (2014). RVboost: RNA-seq variants prioritization using a boosting method. *Bioinformatics* 30, 3414–3416. doi: 10.1093/bioinformatics/btu577

Wolfien, M., Rimmbach, C., Schmitz, U., Jung, J. J., Krebs, S., Steinhoff, G., et al. (2016). TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics* 17:21. doi: 10.1186/s12859-015-0873-9

Zhao, Z., Fu, Y.-X., Hewett-Emmett, D., and Boerwinkle, E. (2003). Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* 312, 207–213. doi: 10.1016/s0378-1119(03)00670-x

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606

Zhou, Y., Lee, J., Reno, C. M., Sun, C., Park, S. W., Chung, J., et al. (2011). Regulation of glucose homeostasis through a XBP-1-FoxO1 interaction. *Nat. Med.* 17, 356–365. doi: 10.1038/nm.2293

Zhuo, Z., Lamont, S. J., and Abasht, B. (2017). RNA seq analyses identify frequent allele specific expression and no evidence of genomic imprinting in specific embryonic tissues of chicken. *Sci. Rep.* 7:11944.

Ensembl genome browser 94 (2020). rs738179721 (SNP) - Explore this variant - Gallus gallus 5 - Archive Ensembl. Available online at: http://oct2018.archive.ensembl.org/Gallus_gallus/Variation/Explore?r=15:7763034-7764034;v=rs738179721;vdb=variation;vf=16501295 (accessed January 5, 2021).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

2.2. Importance de la « phase » entre SNP proches dans la prédiction des effets des variants dans les régions codantes *(Résumé d'article)*

Aparté : Suite à une relecture récente de l'article, une erreur a été identifiée pour la Figure 3. En effet, celle-ci correspond à une version antérieure au processus de relecture par les pairs et les valeurs indiquées ont légèrement fluctuées. Le journal a été contacté pour appliquer la correction. En attendant la modification, la figure corrigée est disponible dans la partie « Valorisation associée » (§2.2.4).

2.2.1. Contexte et objectifs

Les nouvelles technologies de séquençage permettent de détecter à l'échelle du génome entier les micro-variations du génome par rapport à un génome de référence, et ce, notamment grâce au séquençage d'ADN génomique ou en ciblant plus spécifiquement les régions transcrites par l'intermédiaire du séquençage d'exome entier (WES) ou d'ARN (RNAseq) (voir Introduction §2.1 et Résultats §2.1). Parmi les deux types de variations détectés, les polymorphismes mononucléotidiques (SNP) représentent 90% de ces variations contre 10% pour les insertions/délétions (INDEL). Bien que la plupart soient situés dans des régions non codantes (voir Introduction §3.2), les SNP les plus étudiés ont été choisis dans des régions codantes où leurs impacts sur la structure de la protéine sont plus facilement prédictibles [495] spécifiquement par différents logiciels bio-informatiques tels que VEP (*Variant Effect Predictor*), SnpEff ou encore ANNOVAR (*ANNOtate VARiation*) [517, 521, 522]. Cependant, ces outils considèrent chaque SNP individuellement, or chaque individu est caractérisé par ses deux haplotypes parentaux. Ainsi, des SNP présents au sein d'un même codon et formant un unique haplotype chez un individu peuvent exister. Ce groupe de SNP est alors appelé "variants multinucléotidiques" (MNV pour *Multi-Nucleotide Variants*) (voir Figure 1). L'analyse des MNV nécessite donc l'identification correct des haplotypes (phasage) des SNP constitutifs [513–515]. Dans ce cadre, différentes méthodes de phasage des SNP ont été utilisées [523–527] et ont été appliquées à différents ensembles de SNP chez l'humain [389, 495, 524–528], principalement sur la base des exomes.

L'objectif de ce travail est d'évaluer les erreurs de prédiction des effets des SNP sur les protéines dues aux MNV chez la poule à partir de 3,3M de SNP avec génotypes fiables détectés par 744 RNAseq multi-tissus pour 355 individus de 11 populations variées [529]. Par définition, cette étude se limite donc aux SNP phasés dans les codons. Le phasage a été ici observé par

une approche « *read-based* », qui évalue si des variants proches sont supportés par les mêmes *reads*. Cette méthode a été récemment ajoutée par l'outil HaplotypeCaller de GATK qui fournit cette information de phase dans les fichiers *.vcf* [515].

2.2.2. Résultats et discussion

En utilisant les 3,3M de SNP détectés précédemment à partir de données RNAseq, 260 919 SNP uniques ont été identifiés dans les codons de 26 702 transcrits correspondant à 15 835 gènes, soit 81 % des 19 545 PCG connus. Parmi eux, 11 183 SNP, soit 4,3 % des SNP présents dans les codons, étaient regroupés en 5 533 MNV répartis au sein de 4 415 transcrits, soit 2 916 gènes. Environ 98 % (n = 5416) des MNV contenaient 2 SNP avec des proportions similaires concernant les positions des SNP dans le codon. Comme visible dans la Table 1, le nombre de MNV identifiés diminuait lorsque le nombre d'individus les supportant augmentait. En effet, 31 % des 5416 MNV à 2 SNP n'étaient observés que chez un seul individu. Ainsi, afin d'assurer la fiabilité des MNV, seul ceux identifiés pour au moins cinq individus ont été conservées. Cela représente au final 2 965 MNV pour 2 636 transcrits et 1 792 gènes.

En se concentrant uniquement sur les 2 965 MNV présents chez au moins cinq individus, leurs conséquences fonctionnelles ont été comparées avec celles des 5 930 SNP constitutifs (Figure 3). Le plus grand flux de variation d'impact entre MNV et SNP concernait les *stop gained* prédit à l'origine pour les SNP où 95,6 % (87/91) de ces cas ont été re-prédit comme des *missense variant* pour les MNV. Le deuxième et le troisième plus grands flux impliquaient les *missense variant*, pour lesquelles 37,3 % avaient un acide aminé prédit différent (1038/2780) et 3,0 % étaient re-prédit en *synonymous variant* (83/2780). Parmi les 87 *stop gained* « sauvés » observés chez cinq individus, 54 % étaient présents chez au moins 15 individus et représentés en moyenne dans cinq populations. La proportion de MNV *stop_gained* « sauvés » (2,9 %) était du même ordre de grandeur que celle rapportée par le consortium gnomAD chez l'humain avec 1 821 MNV *stop gained* « sauvés » pour un total de 31 575 MNV (5,8 %) [528]. Dans une moindre mesure, notons que 9 *missense variant* ont été re-prédits comme *stop gained*. Finalement, la catégorie *stop gained* a drastiquement décliné de 86 % (de 91 à 13) après la prise en compte des MNV, tandis que la catégorie *synonymous variant* a été multipliée par 2, passant de 79 à 159. Ces différents changements

après la prise en compte des MNV ont un impact majeur sur l'interprétation des variants, et sont donc critiques pour une annotation précise de ces derniers. De manière globale, l'impact fonctionnel des SNP considérés de manière individuelle différait de celui des MNV dans 41,1 % des cas. Ce pourcentage d'erreur de prédiction est cohérent avec les 60 % de ré-annotation des MNV humains récemment rapportés par le consortium gnomAD [528]. De tels résultats montrent l'importance de prêter attention à ces MNV, comme l'ont souligné McLaren et al. [517] : "Les outils d'annotation actuels, y compris VEP, annotent chaque variant d'entrée indépendamment, sans tenir compte des effets composés potentiels de la combinaison d'allèles alternatifs sur plusieurs loci de variants".

Afin d'illustrer l'impact des MNV, un exemple d'erreur de prédiction d'un codon stop (*stop gained*) est présenté pour le gène SLC27A4 et pour la lignée de poule FLLL (Figure 4). Deux SNP, rs316701182 et rs15031398, déjà rapportés dans la base de donnée dbSNP d'Ensembl [508], ont été prédits comme variant *stop gained* (TGA) et *synonymous variant* (CGC) respectivement. Ces SNP étaient présents dans la population FLLL (n = 24) avec des fréquences > 20% et avec des contrastes entre les lignées FL (n = 12) et LL (n = 12) sélectionnées de façon divergente pour le poids de tissu adipeux [510, 511]. Alors que pour la lignée FL, le SNP rs15031398 apparaissait absent, pour la lignée LL, aucun animal ne portait l'haplotype prédit *stop gained*. En effet, le rs316701182 était toujours en phase avec le rs15031398 sous la forme d'un MNV sauvant le *stop gained* (Figure 4C). L'absence de l'haplotype *stop gained* est cohérent avec des études rapportant une létalité de l'inactivation de SLC27A4 chez la souris [530–534]. Comme visible dans la Figure 4D, ce gène est fortement exprimé (> 10 TPM) dans le foie, les ovaires, la peau et l'intestin chez la poule, ce qui est cohérent avec son rôle connu dans le transport des acides gras à longue chaîne, fortement exprimée dans divers tissus dont l'intestin [535, 536]. L'haplotype re-prédit pourrait cependant être impliqué dans le phénotype maigre de la lignée LL car la conséquence est considérée comme ayant un impact sévère sur la fonction du gène par SIFT et est fréquemment observée (42 %). De plus, cet haplotype est absent dans la lignée FL. Des résultats similaires ont été obtenus en séquençant le gène après PCR chez 58 poules (29 FL et 29 LL) cependant le lien de causalité entre l'haplotype re-prédit et une faible adiposité chez la lignée LL doit être confirmé par des études d'associations génétiques.

2.2.3. Matériels et démarches

L'annotation des effets des SNP sur la structure protéique a été réalisée avec VEP [517]. Les MNV ont été identifiés par un script développé au laboratoire et suivant trois étapes (voir Figure 2) :

1) Détection des SNP dans un même codon via la conception d'identifiants uniques (*e.g*, *ENSGALT0000125_12/144)* résultant de l'agrégation de l'identifiant du transcrit (*ENSGALTxx*) et du numéro du codon affecté (*aa/bb*) ;

2) Sélection des SNP phasés et co-localisés au sein du même codon (MNV) grâce aux informations haplotypiques fournies par HaplotypeCaller de GATK [513–515] ;

3) Re-prédiction des conséquences des MNV sur la structure protéique en suivant la même stratégie d'annotation que VEP et comparaison aux effets des SNP constitutifs du MNV calculés séparément.

Afin de comparer les conséquences prédites d'une part pour les MNV et d'autre part pour les SNP considérés indépendamment, nous avons sélectionné uniquement la conséquence la plus importante par codon pour les SNP constitutifs en utilisant l'ordre de priorité suivant, allant des conséquences les plus graves aux plus faibles :

1) Apparition d'un codon stop (*stop gained*) ;
2) Perte d'un codon stop (*stop lost*) ;
3) Perte d'un codon start (*start lost*) ;
4) Mutation faux-sens changeant l'acide aminé (*missense variant*) ;
5) Modification du codon stop mais maintien de celui-ci (*stop retained*) ;
6) Mutation synonyme changeant le codon mais ne modifiant pas l'acide aminé (*synonymous variant*).

Pour les MNV induisant des substitutions d'acides aminés (*missense variant*), deux cas ont été distingués : soit *i)* l'acide aminé est identique à celui de l'un des SNP constituants, soit *ii)* l'acide aminé est différent.

2.2.4. Valorisation associée

Ces travaux ont fait l'objet :

- d'une publication : **Degalez F***, Jehl F*, Muret K, Bernard M, Lecerf F, Lagoutte L, Désert C, Pitel F, Klopp C and Lagarrigue S (2021). Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained. Frontiers in Genetics. doi: 10.3389/fgene.2021.659287. ***Cet article est reproduit ci-après ;***

Figure 3 corrigée

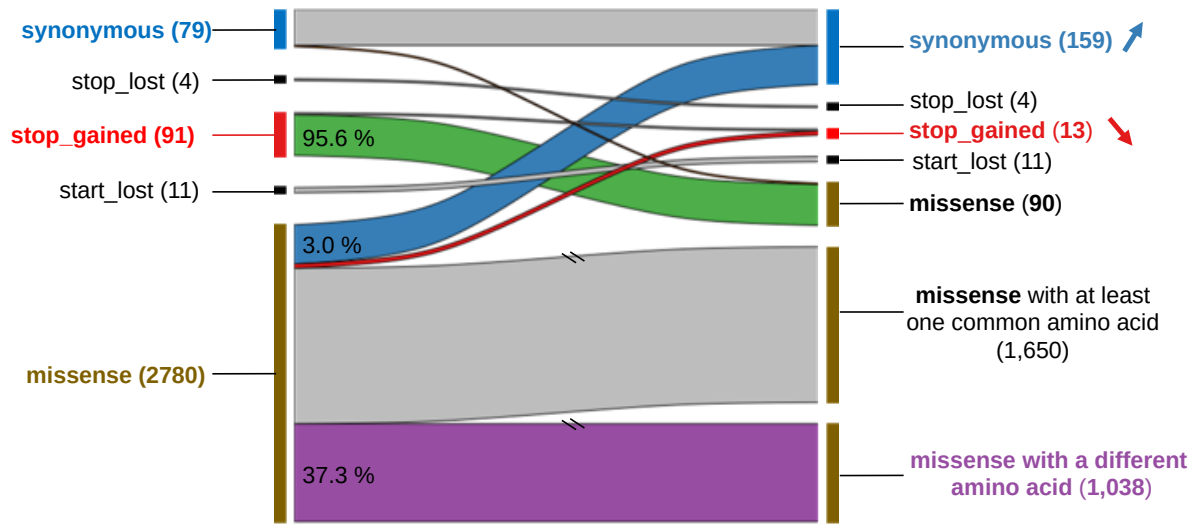


FIGURE 3 | Comparison of the functional impact of MNVs (right) and their component SNPs (left) for each of the 2,965 MNVs. Left: The consequence originally predicted for the component SNPs, the most severe impact being retained by the codon (see **Figure 2** or the text for the order). Right: The new prediction associated with the MNVs. For each category of functional predictions of the component SNPs (left), the numbers and percentages are given with new predictions due to the associated MNV. The two slashes indicate that the scale has been adapted (reduction by five times) for better readability.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 July 2021
doi: 10.3389/fgene.2021.659287

Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained

Fabien Degalez[1†], Frédéric Jehl[1†], Kévin Muret[1], Maria Bernard[2,3], Frédéric Lecerf[1], Laetitia Lagoutte[1], Colette Désert[1], Frédérique Pitel[4], Christophe Klopp[2] and Sandrine Lagarrigue[1]*

[1] INRAE, INSTITUT AGRO, PEGASE UMR 1348, Saint-Gilles, France, [2] INRAE, SIGENAE, Genotoul Bioinfo MIAT, Castanet-Tolosan, France, [3] INRAE, AgroParisTech, Université Paris-Saclay, GABI UMR 1313, Jouy-en-Josas, France, [4] INRAE, INPT, ENVT, Université de Toulouse, GenPhySE UMR 1388, Castanet-Tolosan, France

OPEN ACCESS

Edited by:
Hans Cheng,
Agricultural Research Service (USDA), United States

Reviewed by:
John B. Cole,
United States Department of Agriculture (USDA), United States
Robert D. Schnabel,
University of Missouri, United States

***Correspondence:**
Sandrine Lagarrigue
sandrine.lagarrigue@agrocampus-ouest.fr

[†]These authors share first authorship

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 27 January 2021
Accepted: 27 May 2021
Published: 07 July 2021

Citation:
Degalez F, Jehl F, Muret K, Bernard M, Lecerf F, Lagoutte L, Désert C, Pitel F, Klopp C and Lagarrigue S (2021) Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained. Front. Genet. 12:659287. doi: 10.3389/fgene.2021.659287

Most single-nucleotide polymorphisms (SNPs) are located in non-coding regions, but the fraction usually studied is harbored in protein-coding regions because potential impacts on proteins are relatively easy to predict by popular tools such as the Variant Effect Predictor. These tools annotate variants independently without considering the potential effect of grouped or haplotypic variations, often called "multi-nucleotide variants" (MNVs). Here, we used a large RNA-seq dataset to survey MNVs, comprising 382 chicken samples originating from 11 populations analyzed in the companion paper in which 9.5M SNPs— including 3.3M SNPs with reliable genotypes—were detected. We focused our study on in-codon MNVs and evaluate their potential mis-annotation. Using GATK HaplotypeCaller read-based phasing results, we identified 2,965 MNVs observed in at least five individuals located in 1,792 genes. We found 41.1% of them showing a novel impact when compared to the effect of their constituent SNPs analyzed separately. The biggest impact variation flux concerns the originally annotated stop-gained consequences, for which around 95% were rescued; this flux is followed by the missense consequences for which 37% were reannotated with a different amino acid. We then present in more depth the rescued stop-gained MNVs and give an illustration in the *SLC27A4* gene. As previously shown in human datasets, our results in chicken demonstrate the value of haplotype-aware variant annotation, and the interest to consider MNVs in the coding region, particularly when searching for severe functional consequence such as stop-gained variants.

Keywords: MNV, SNP, variation, rescued stop-gained, *SLC27A4*, FATP4

INTRODUCTION

Next-generation sequencing has given access to genomes at the nucleotide level through DNA-seq but also specifically to expressed regions by whole-exome sequencing (WES, originally focusing on exonic parts of the genome) or RNA-seq. These data enable us to call genetic variations by spotting differences between aligned reads and the species reference genome or among aligned

reads. Among these genetic variations, single-nucleotide polymorphisms (SNPs) are the most frequent and most studied variations. Although most variations are located in non-coding regions, the most analyzed lie in protein-coding regions where their potential impact(s) on the protein are relatively easy to predict. For example, in a study using 60,706 human exomes, the Exome Aggregation Consortium (ExAC) identified 3,230 genes with near-complete depletion of predicted protein-truncating variants. Of these genes, 72% have not been related to any known human disease phenotype (Lek et al., 2016). Different popular tools have been developed this last decade to predict SNPs' effects on proteins such as *Variant Effect Predictor* (VEP) (McLaren et al., 2016), SnpEff (Cingolani et al., 2012), or *ANNOtate VARiation* (ANNOVAR) (Wang et al., 2010). But these tools consider each variation location individually, as if it they were specific to "reference" nucleotides. However, SNPs can be grouped by two or more coexisting variants present in the same haplotype (in the same individual), in which case they are called "multi-nucleotide variants" (MNVs). An example of MNV in one individual (with two nearby SNPs) is given in **Figure 1**. When such MNVs occur within a codon, the amino acid modification caused by this MNV may be different from protein change resulting from each constituent SNPs taken individually, leading to a risk of erroneous functional consequence prediction, as depicted in **Figure 1**. MNV identification tools have been developed using different methods for phasing SNPs [MAC (Wei et al., 2015), varDic (Lai et al., 2016), COPE (Cheng et al., 2017), BCFtools (Danecek and McCarthy, 2017), and MACARON (Khan et al., 2018)] and have been applied to different human genetic variant datasets [1,000 Genomes Project dataset (Cheng et al., 2017; Danecek and McCarthy, 2017; Khan et al., 2018; Wang et al., 2020), ExAC (Lek et al., 2016), The Cancer Genome Atlas (Lai et al., 2016), or gnomAD consortium (Wang et al., 2020)], mainly based on exomes.

To our knowledge, no study has been conducted on MNVs in livestock species. The aim of this paper is to focus on MNVs occurring in protein-coding regions to provide examples and evaluate the functional consequences of resulting mis-annotations. Considering this aim, we used 9.5M SNPs recently detected in 382 chickens from 767 multi-tissue RNA-seq, enriched by construction in expressed regions and therefore in protein-coding regions. From this 9.5M SNPs, we focused on the 3.3M SNPs with reliable genotypes [see the companion paper (Jehl et al., 2021)]. MNV identification requires properly phased variants, i.e., to be located either on the same haplotype (called therefore MNV) or on two different haplotypes (a case of individual SNPs) (see **Figures 1A,B**). Different SNP phasing strategies exist: (*i*) population-based phasing, using statistical inference of phase from haplotypes shared among individuals of a large genotyped population; (*ii*) family-based phasing, which analyzes the co-transmission of variants between parents and offsprings; and (*iii*) read-based phasing, which evaluates whether close variants are present on the same reads in the DNA-seq or RNA-seq data. Read-based phasing is particularly relevant for close variants, making this method appropriate for MNV analysis in codons, in which variants fall within a maximum distance of 2 bp from one another. Therefore, in this study, we have chosen to identify MNVs by using the read-based phasing provided by the HaplotypeCaller tool of the GATK toolkit, in the VCF file through additional fields (**Figure 1C**, with the PID and PGT fields) recently added by the common variant caller (DePristo et al., 2011; Van der Auwera et al., 2013; McKenna et al., 2020).

MATERIALS AND METHODS

SNP Dataset

The 3,276,615 SNPs analyzed in this study have been detected following the method presented in the companion paper (Jehl et al., 2021) using 767 multi-tissue RNA-seq of 382 birds from 11 chicken populations (see **Additional File 1**). This SNP set corresponds to the union of the SNPs with reliable genotypes found in each population (list available on http://www.fragencode.org/lnchickenatlas.html). Briefly, variant detection was performed for each sample using the HaplotypeCaller tool of GATK toolkit (DePristo et al., 2011; Van der Auwera et al., 2013; McKenna et al., 2020) 3.7.0 with options "--stand_call_conf 20.0," "--min_base_quality_score 10," and "--min_mapping_quality_score 20" (which are the defaults values). The "GenotypeGVCFs" function was then used with the option "--stand_call_conf 20.0," to jointly genotype all these samples into one VCF per tissue. First, biallelic SNPs were then extracted using the "SelectVariant" function with option "--selectType SNP –restrictAllelesTo BIALLELIC." Variants were filtered using "VariantFiltration" with "QD $<$ 2" and "FS $>$ 30." Considering genotypes, variants were selected with a "(5.reads.DP) genotype CR $\geq$ 20%" and a "CR $\geq$ 50%." The 11 populations include a red jungle fowl population (RJFh), three broiler populations with one experimental line (FLLL) and two commercial ones (Cobb, HerX), six layer populations with two brown-egg commercial lines (Novo1 and Novo2), two brown-egg experimental populations (RpRm and LSnu), two white-egg or cream-egg experimental populations (FrAg and FAyo), and finally a cross between white- and brown-egg experimental lines (Rmx6).

Analysis of the Functional Impact of Each Individual SNP in the Coding Regions

VEP v92 (McLaren et al., 2016) with a GTF file enriched in long non-coding genes ("--gtf") was used for effect prediction of each SNP, with "--everything" and "--total_length" options to respectively, obtain SIFT score predictions and lengths of cDNA, CDS, and protein positions (Ng and Henikoff, 2003; Sim et al., 2012).

MNV Calling and Recalculation of Consequences

The script to detect the MNV and to calculate the consequences is available in **Additional File 2**.

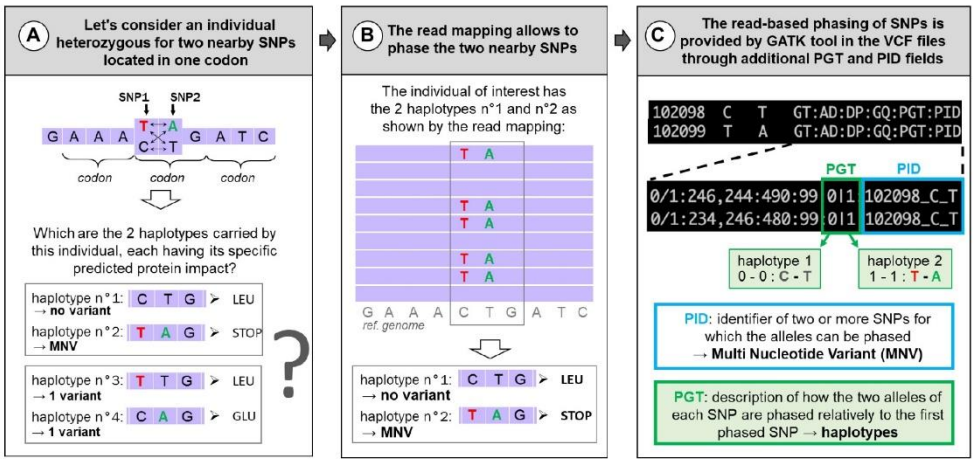


FIGURE 1 | Example of MNVs: predicted impact on the associated protein **(A)** and how to identify them **(B,C)**. **(A)** Example of an MNV composed of two nearby SNPs in one codon and its four potential haplotypes in the population and their predicted impact on the associated protein. In contrast to other haplotypes, haplotype no. 2 contains two variants (T and A) and corresponds to an MNV. **(B)** The IGV (Integrated Genome Browser; Robinson et al., 2011) screen shot indicates the principle of read-based phasing of SNPs: short read mapping against the reference genome of the heterozygous individual allows us to phase both SNPs giving two haplotypes: C with T (reference alleles) on one side and T with A (alternative alleles) on the other side. When translated, these two haplotypes correspond to a leucine or a STOP codon and not to a simple amino acid change (LEU → GLU) if the two haplotypes had been composed by only one reference and one variant as shown in **(A)**. **(C)** Information (PID and PGT) provided by GATK in the VCF files about the phased SNPs according to the read-based phasing shown in **(B)**.

Detection of SNPs Located in the Same Codon

With the information produced by VEP, an ID composed of the "transcriptID" and the "position of the SNP in the coding sequence (expressed in codon number)" was created for each consequence. Through this approach, the same codon of the same transcript supporting at least two different SNPs will have the same ID. Thus, only duplicated IDs were kept as they correspond to those containing two or more SNPs.

Detection of Co-located SNPs Carried by the Same Haplotype (MNV)

To test if the SNPs located in the same codon were also present in the same haplotype, the VEP file generated in the previous step and the VCF file were joined by the "SNPid" key, equivalent to "CHR_POS_REF/ALT." The resulting file (VEP merged to VCF information) contained SNPs on the same codon with additional information about their phase (PID and PGT). Finally, the SNPs which were phased (i.e., same PID) and co-located in a codon were extracted: they correspond to MNVs containing two or three phased SNPs in the same codon.

Recalculation of the Consequences

With the R package Biostrings v2.50.2 (Pagès et al., 2021), the associated amino acids were produced for each MNV, and with the same strategy as VEP being adopted, the MNV consequences were established.

Analysis of MNV Functional Impacts and Comparison With the Constituent SNP Impacts

To compare MNV and independent SNP consequences, we selected only the most impactful consequence per codon for these constituent SNPs using the following order of priority from severe to weak consequences: (1) stop-gained, (2) stop-lost, (3) start-lost, (4) missense_variant, (5) stop-retained_variant, and (6) synonymous_variant.

For MNVs with a missense annotation corresponding to a missense annotation for both constituent SNPs, we distinguished two cases:

- missense MNV with an amino acid different from those predicted by the constituent SNPs (SNP1: Missense A; SNP2: Missense B → MNV: Missense C) and
- missense MNV with an amino acid common to one of two amino acids predicted by the constituent SNPs (SNP1: Missense A; SNP2: Missense B → MNV: Missense A or B).

In order to visualize the results, we produced an alluvial plot using the R "alluvial" package v0.1-2 (Bojanowski, 2020).

GO or KEGG Term Enrichment Analysis

The enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) terms in the gene set of interest was performed using the STRING v11.0. tool (Szklarczyk et al., 2019), and a GO or KEGG term was found significantly enriched if the BH-adjusted $p \leq 5\%$.

DNA Sequencing of *SLC27A4*

Five microliters of DNA samples was mixed with 5 µl of GoTaq Flexi Buffer 5 × , 2 µl of $MgCl_2$ solution (25 mM), 0.125 µl of GoTaq DNA polymerase (5 U/µl) (Promega, catalog number: M891), 0.5 µl of dNTPs 10 mM, 12.5 µl H_2O, and 1.25 µl of specific reverse (CATTCCCGTAGTGCCAGAGG) and forward primers (GCACTTTCTGGTGCAAAGCA) at 10 µM. Reaction mixtures were then incubated in a T100 thermal cycler (Bio-Rad, Marne la Coquette, France) for 30 cycles with 30 s at 94°C, 30 s at 60°C, and 30 s at 72°C. The amplification products were then deposited on a 2% agarose gel and sent for sequencing (Genoscreen) to verify their location on the chicken genome.

RESULTS AND DISCUSSION

Read-Based Phasing for Identification of MNVs

Using 3.3M SNPs previously detected from 767 multi-tissue RNA-seq of 382 animals from 11 chicken populations and therefore enriched in coding regions [see the companion paper (Jehl et al., 2021), section "Materials and Methods"], we identified 260,919 unique SNPs in 26,702 transcripts corresponding to 15,835 genes out of 19,545 protein-coding genes (**Figure 2**, right part—in yellow).

As shown in **Figure 2** (left part—in green), we then defined an MNV in a codon as a group of two or three phased SNPs, i.e., existing on the same haplotype in the same individual. We found 11,183 SNPs (4.3% of the SNPs in codons) as constituent variants of 5,533 MNVs, which corresponded to 4,415 transcripts and 2,916 genes. Most of them (98%: 5,416) contained two SNPs with similar proportions (1/3) by constituent SNP position in the codon (1–3, 1–2, and 2–3, **Figure 2**, left). In order to ensure the reliability of the MNVs, we selected MNVs observed in at least five individuals. Out of the 5,416 MNVs with two SNPs, 2,965 MNVs were present in at least five individuals, corresponding to 2,636 transcripts and 1,792 genes. No GO terms or KEGG terms were found as significantly enriched for this gene list, suggesting that no specific biological pathway was impacted by MNVs. **Table 1** gives the distribution of MNVs and their consequences according to the individual number supporting the MNV (ranging from 2 to 100 individuals). We can note that 31% of the 5,416 MNVs with two SNPs are observed only in a single individual and are here considered as erroneous, likely due to sequencing errors.

Functional Impact Comparison of MNVs and Their Component SNPs

Focusing on the 2,965 MNVs present in at least five individuals, we then compared their functional consequences with those of the 5,930 constituent SNPs as illustrated by the right part of the workflow provided in **Figure 2**. For such a comparison, we retained for each MNVs, the most severe consequence of the constituent SNPs according to the order indicated in **Figure 2** (bottom right). The alluvial plot in **Figure 3** depicts the consequence variations before (left) and after (right) taking the MNV impacts into account, according to the different consequence categories; the details of impact variation per MNV are given in **Additional File 3** for the whole 2,965 MNV set. We can observe in **Figure 3** that the biggest change in variant impacts concerns the originally stop-gained consequence categories, for which 95.6% were re-predicted as missense (green flux: 87 out of the 91 stop-gained initially predicted). The second and third biggest fluxes concern missense consequence categories, for which 37.3% had a different predicted amino acid (violet flux: 1,038 MNVs out of the 2,780 initial missenses), and 3.0% became synonymous variants (blue flux: 83 MNVs out of the initial missenses). The distribution of re-prediction fluxes is provided in **Table 1** as a function of the individual number supporting the MNV among the 382 individuals analyzed. Among the 87 rescued stop-gained observed in five individuals, half (47) are observed in at least 15 individuals and are present on average in five populations (see **Additional File 4**). Out of the MNVs, the proportion of rescued stop-gained MNVs (2.9%), defined as at least one of the individual SNPs creating a nonsense mutation but not the resulting MNV, is in the same order of magnitude as the one reported by the gnomAD consortium with 1,821 rescued stop-gained MNVs out of 31,575 human MNVs (5.8%) (Wang et al., 2020). Genes with a stop-gained MNV rescued in the missense variant are available in **Additional File 4** with the population affected and the individual number per population carrying these MNVs. To a lesser extent, nine missenses were re-predicted as stop-gained, which would have gone unnoticed without re-prediction. After a deeper investigation with the IGV browser, these re-predicted stop-gained variants seem to be present since they were not located in a potential exon skipping. Finally, this stop-gained category drastically declined by 86% (from 91 to 13) after considering MNVs, whereas the synonymous category was increased by twofold (from 79 to 159). These different category changes after considering MNVs have a major impact on variant interpretation and thus are critical for accurate variant annotation. More broadly, when the MNVs were considered together, the resulting functional impact differed from the independent impacts of the individual variants in 41.1% of the analyzed MNVs. This large percentage of mis-annotations is relatively consistent with ∼60% of reannotations in human MNVs recently reported by the gnomAD consortium in coding regions (Wang et al., 2020). Such results show the importance of paying attention to these MNVs as highlighted by McLaren et al. (2016): "Current annotation tools, including the VEP, annotate each input variant independently, without considering the potential compound effects of combining alternate alleles across multiple variant loci."

Example of an Erroneously Predicted Stop-Gained

As an example of erroneously predicted stop-gained, we present the case of the *SLC27A4* gene, which is located on the reverse strand of chicken chromosome 17 (ENSGALG00000004965) (**Figure 4A**). In this gene, two SNPs rs316701182 and rs15031398, already reported in the Ensembl SNP database (Ensembl, 2018), were respectively, predicted as a stop-gained variant

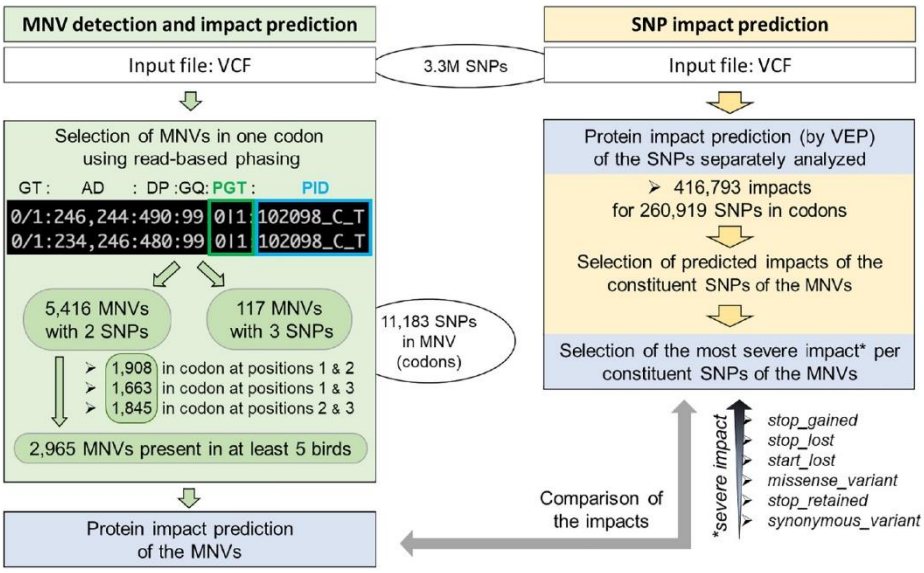


FIGURE 2 | Workflow of MNV detection in coding regions and functional consequence prediction. Left: MNV detection from 3.3M SNPs previously identified using RNA-seq of 382 chickens (companion article; Jehl et al., 2021). Right: MNV constituent SNP selection and protein impact selection of these SNPs separately analyzed by VEP.

TABLE 1 | Occurrences for each type of re-prediction according to the number of individuals carrying the MNV.

SNP annotation	→	MNV annotation	Number of individuals carrying the MNV										
			1	**2**	**3**	**4**	**5**	**10**	**15**	**20**	**30**	**50**	**100**
	Eased impact												
Missense	→	Synonymous	110	91	86	86	**83**	76	74	74	72	66	54
Stop_gained	→	Missense	194	118	102	99	**87**	63	47	39	33	27	14
	Equal impact												
Stop_lost	→	Stop_lost	8	4	4	4	**4**	2	2	2	2	1	0
Start_lost	→	Start_lost	19	11	11	11	**11**	8	7	6	5	3	1
Synonymous	→	Synonymous	95	84	81	78	**76**	69	63	58	54	45	32
Missense	→	Missense	4,932	3,425	3,107	2,854	**2,688**	2,162	1,876	1,652	1,369	1,083	716
Stop_gained	→	Stop_gained	12	6	5	5	**4**	3	2	2	2	1	0
	Aggravated impact												
Stop_retained	→	Stop_lost	1	0	0	0	**0**	0	0	0	0	0	0
Synonymous	→	Missense	15	6	6	4	**3**	2	1	1	1	1	1
Missense	→	Stop_gained	30	13	11	10	**9**	6	4	3	3	2	1
		Total MNVs	5,416	3,758	3,413	3,151	**2,965**	2,391	2,076	1,837	1,541	1,229	819

*The column in bold corresponds to MNVs observed in at least five individuals (MNV set used in **Figure 3**).*

(TGA; stop-gained) and a synonymous variant (CGC; arginine) when compared to the reference haplotype (CGA; arginine) (**Figure 4B**). These SNPs were present in the FLLL population with frequencies > 20% and interestingly with contrasted frequencies between FL (fat line) and LL (lean line), two subpopulations divergently selected for adipose tissue weight (Leclercq et al., 1980). The rs15031398 SNP is absent in FL (**Figure 4B**); in the LL population in which we observed both SNPs (**Figure 4C**), we did not find any animal with the TGA (stop-gained) haplotype (composed of one variant only), with the rs316701182 T variant being always associated with the rs15031398 C variant within the "TGC" MNV. The absence of TGA (stop-gained) haplotype is consistent with several *SLC27A4*-knockout mouse studies which report prenatal lethality (Gimeno et al., 2003) or neonatal lethality (Herrmann et al., 2003; Moulson et al., 2003; Lin et al., 2010; Tao et al., 2012). The *SLC27A4* gene codes fatty acid transport protein 4 (FATP4), which is particularly involved in the uptake of long-chain fatty acids

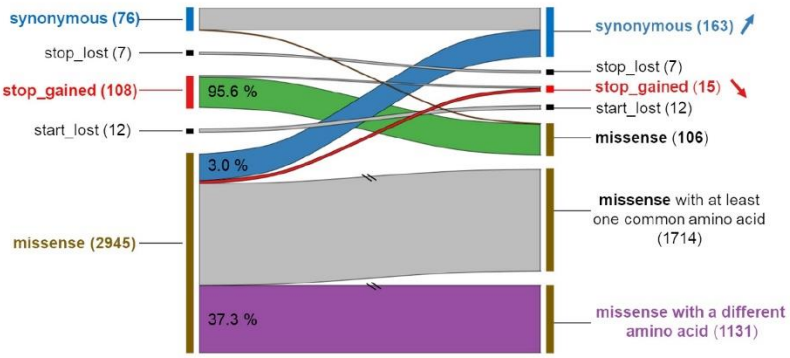
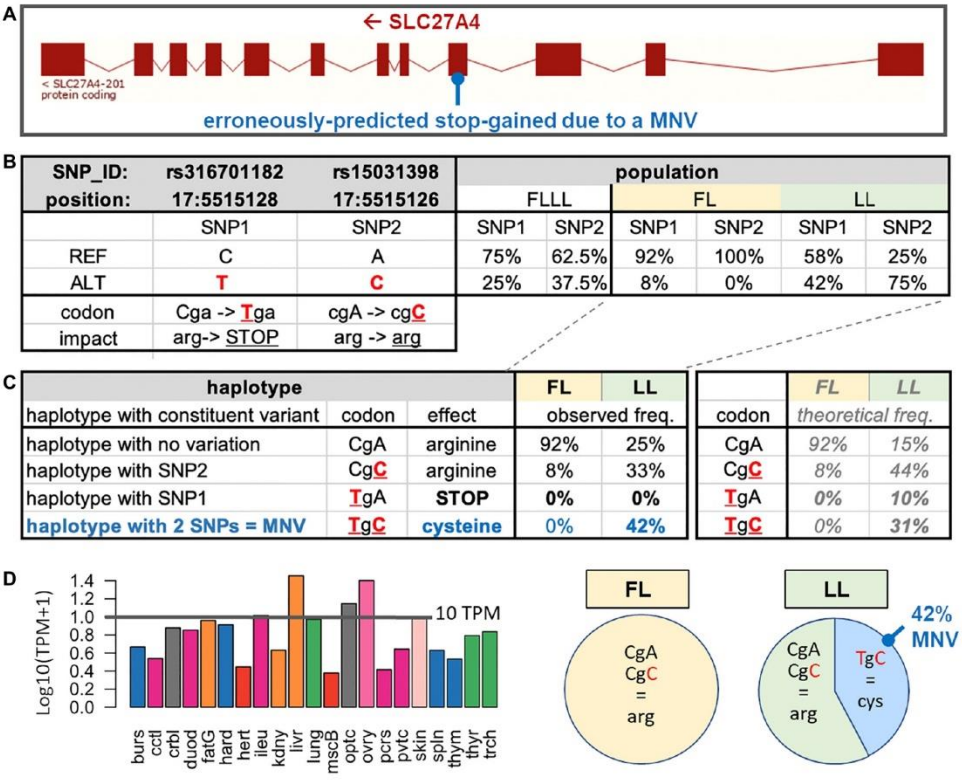


FIGURE 3 | Comparison of the functional impact of MNVs (right) and their component SNPs (left) for each of the 2,965 MNVs. Left: The consequence originally predicted for the component SNPs, the most severe impact being retained by the codon (see **Figure 2** or the text for the order). Right: The new prediction associated with the MNVs. For each category of functional predictions of the component SNPs (left), the numbers and percentages are given with new predictions due to the associated MNV. The two slashes indicate that the scale has been adapted (reduction by five times) for better readability.



SNP_ID:	rs316701182	rs15031398	population					
position:	17:5515128	17:5515126	FLLL		FL		LL	
	SNP1	SNP2	SNP1	SNP2	SNP1	SNP2	SNP1	SNP2
REF	C	A	75%	62.5%	92%	100%	58%	25%
ALT	T	C	25%	37.5%	8%	0%	42%	75%
codon	Cga -> Tga	cgA -> cgC						
impact	arg-> STOP	arg -> arg						

haplotype			FL	LL
haplotype with constituent variant	codon	effect	observed freq.	
haplotype with no variation	CgA	arginine	92%	25%
haplotype with SNP2	CgC	arginine	8%	33%
haplotype with SNP1	TgA	**STOP**	**0%**	**0%**
haplotype with 2 SNPs = MNV	TgC	**cysteine**	0%	**42%**

	FL	LL
codon	*theoretical freq.*	
CgA	*92%*	*15%*
CgC	*8%*	*44%*
TgA	*0%*	*10%*
TgC	*0%*	*31%*

FIGURE 4 | *SLC27A4* with an MNV composed of two phased SNPs observed in the experimental divergent lean line (LL). **(A)** Exon structure of the *SLC27A4* gene and the MNV location. **(B)** For the two SNPs (SNP1: rs316701182 and SNP2: rs15031398) related to the MNV of interest, the allele position (on the Galgal5 genome), functional impact on the associated protein, and frequencies in the FLLL population are indicated, and the two FL and LL subpopulations are divergently selected on abdominal fat weight. **(C)** Effects of the four haplotypes related to SNP1 and SNP2 separately analyzed by VEP and frequencies in LL (n = 12) and FL (n = 12) subpopulations and focus on the percentage of observed haplotypes in the two FL and LL subpopulations. The haplotypes were determined through the IGV browser of mapped RNA-seq reads against the chicken genome. **(D)** Tissue expression of the gene in a chicken RNA-seq dataset composed of 21 tissues (Jehl et al., 2020).

(LCFAs); this gene is highly expressed in various chicken tissues as shown in **Figure 4D** with an expression > 10 TPM in the liver, ovary, optical system, skin, and intestine (ileum). Interestingly, FATP4 is thought to play a major role in dietary fatty acid uptake in intestinal epithelial cells (Hirsch et al., 1998) and in physiological uptake across cell membranes of LCFAs, which are key metabolites for energy generation and storage; it is viewed as a target to prevent or reverse obesity (Hirsch et al., 1998; Schaffer, 2002). FATP4 could be then related to the lean phenotype of the LL population for two reasons. First, the "TGC" (cysteine) MNV haplotype is reported as a severe change by the SIFT software package compared to the reference "CGA" (arginine) haplotype, suggesting a severe impact on the FATP4 protein function. Second, this "TGC" MNV haplotype is absent in FL birds, whereas it is frequent (42%) in 12 LL birds, with a higher frequency than expected (**Figure 4C**). We confirmed these results by extending this analysis to 58 birds (29 birds per line) using PCR amplification of the region of interest followed by Sanger sequencing. No rs15031398 was identified in the FL line. In the LL line, we observed 12 birds carrying the "TGC" MNV haplotype (three homozygous and nine heterozygous) and no bird with the TGA (stop-gained) haplotype. These results suggest a strong but not lethal impact of the MNV haplotype on the FATP4 protein function, which could then participate to the lean phenotype of the LL line. However, a genetic association study is needed to support a potential causal link between the FATP4 dysfunctional MNV and a low adiposity in the LL line compared to the FL line.

CONCLUSION

We have shown that MNVs represent an important class of genetic variations since they have a significant impact on polymorphism functional interpretation with roughly 40% of MNVs in our dataset inducing reannotation. These reannotations show a decreased impact severity of MNVs when compared to their constituent SNPs, at least for the stop-gained category. As previously demonstrated in human studies, our results in chicken demonstrate the value of haplotype-aware variant annotation and the interest to consider MNVs in coding region particularly when focusing on severe functional consequences such as stop-gained. We illustrated such a case with an erroneous stop-gained annotation found in the chicken *SLC27A4* gene.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

FD, FJ, and SL conceived the study and drafted the manuscript. FD, FJ, FL, CK, and SL participated to in bioinformatics processing of the RNA-seq data and bioinformatic analyses. LL, CD, and KM participated in the IGV analysis, PCR amplification, and Sanger sequencing for *SLC27A4* analyses. KM, FP, FL, and CK helped to improve the manuscript. All authors read and approved the final version.

FUNDING

The sample and/or data were collected in the frame of projects that received financial support from the European Union's H2020 Program under Grant Agreement No. 633531 (Feed-a-Gene project), the French National Agency of Research (FatInteger project ANR-11-SVS7; ChickStress project, ANR-13-ADAP; EpiBird ANR project PCS-09-GENM-010), and the French institutions Institut Agro (AGROCAMPUS OUEST) and INRAE [Fr-AgENCODE project (2015–2017) and ELASTICe project (2012)]. FJ, FD, and KM were Ph.D. fellows supported by the Brittany region (France) and the INRAE Animal Genetics division. These funding bodies had no role in the design of the study; in the collection, analysis, and interpretation of data; or in the writing of the manuscript.

ACKNOWLEDGMENTS

We are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi: 10.15454/1.5572369328961167E1) for providing help and/or computing and/or storage resources. We thank to Morgane Boutin who passed away this year and who did a lot in *SLC27A4* analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.659287/full#supplementary-material

Additional File 1 | Detail per population of the number of birds, tissues, and samples. The median number of sequenced fragments (× 1E6) is also indicated for each population.

Additional File 2 | Script used for recalculating the consequences for MNVs made up of two SNPs. The script is not generalized and therefore requires adaptations from the users. It is written in bash and R; the change of language is clearly indicated.

Additional File 3 | List of MNVs with two phased SNPs (position on the Galgal5 chicken genome) identified among the 382 chickens from 11 populations with their functional impacts (newConseq) and their constituent SNPs (oldConseq). **(A)** List of 2,965 MNVs identified in at least five individuals. **(B)** List of 5,416 MNVs without filtering on the number of individuals supporting the MNV. For the two tables, the population number (popNb), the total number of individuals (indNb_total), and the population name with the individual number per population (popName_indNb), in which the MNVs have been observed, are provided. AA, amino acid.

Additional File 4 | List of genes presenting a rescued stop-gained according to the number of individuals carrying the MNV. **(A)** Summary of the number of genes detected as "rescued" according to the number of individuals supporting the MNV. **(B)** For each gene, the population number (popNb) and the population name(s) with the individual number per population (popName:indNb) where the MNV is observed are indicated. The column in blue corresponds to MNVs observed in at least five individuals, the MNV set used in **Figure 3**.

REFERENCES

Bojanowski, M. (2020). *mbojan/alluvial. R.* Available online at: https://github.com/mbojan/alluvial. (accessed 8 Jan 2021).

Cheng, S.-J., Shi, F.-Y., Liu, H., Ding, Y., Jiang, S., Liang, N., et al. (2017). Accurately annotate compound effects of genetic variants using a context-sensitive framework. *Nucleic Acids Res.* 45:e82. doi: 10.1093/nar/gkx041

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi: 10.4161/fly.19695

Danecek, P., and McCarthy, S. A. (2017). BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* 33, 2037–2039. doi: 10.1093/bioinformatics/btx100

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498.

Ensembl. (2018). *dbSNP - Gallus Gallus 5 - V94.* ftp://ftp.ensembl.org/pub/release-94/variation/vcf/gallus_gallus (accessed December 11, 2020).*

Gimeno, R. E., Hirsch, D. J., Punreddy, S., Sun, Y., Ortegon, A. M., Wu, H., et al. (2003). Targeted deletion of fatty acid transport protein-4 results in early embryonic lethality. *J. Biol. Chem.* 278, 49512–49516. doi: 10.1074/jbc.m309759200

Herrmann, T., van der Hoeven, F., Gröne, H.-J., Stewart, A. F., Langbein, L., Kaiser, I., et al. (2003). Mice with targeted disruption of the fatty acid transport protein 4 (Fatp 4, Slc27a4) gene show features of lethal restrictive dermopathy. *J. Cell Biol.* 161, 1105–1115. doi: 10.1083/jcb.200207080

Hirsch, D., Stahl, A., and Lodish, H. F. (1998). A family of fatty acid transporters conserved from mycobacterium to man. *Proc Natl Acad Sci U S A.* 95, 8625–8629. doi: 10.1073/pnas.95.15.8625

Jehl, F., Degalez, F., Bernard, M., Lecerf, F., Lagoutte, L., Désert, C., et al. (2021). RNA-Seq data for reliable SNP detection and genotype calling: Interest for coding variant characterization and cis-regulation analysis by allele-specific expression in livestock species. *Front. Genet.* 12:655707. doi: 10.3389/fgene.2021.655707

Jehl, F., Muret, K., Bernard, M., Boutin, M., Lagoutte, L., Désert, C., et al. (2020). An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci. Rep.* 10:20457.

Khan, W., Varma Saripella, G., Ludwig, T., Cuppens, T., Thibord, F., Génin, E., et al. (2018). MACARON: a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data. *Bioinformatics* 34, 3396–3398. doi: 10.1093/bioinformatics/bty382

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., et al. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 44:e108. doi: 10.1093/nar/gkw227

Leclercq, B., Blum, J. C., and Boyer, J. P. (1980). Selecting broilers for low or high abdominal fat: Initial observations. *Br. Poult. Sci.* 21, 107–113. doi: 10.1080/00071668008416644

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291.

Lin, M.-H., Chang, K.-W., Lin, S.-C., and Miner, J. H. (2010). Epidermal hyperproliferation in mice lacking fatty acid transport protein 4 (FATP4) involves ectopic EGF receptor and STAT3 signaling. *Dev. Biol.* 344, 707–719. doi: 10.1016/j.ydbio.2010.05.503

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2020). *The Genome Analysis Toolkit: A Mapreduce Framework For Analyzing Next-Generation Dna Sequencing Data.* Available online at: https://genome.cshlp.org/content/20/9/1297.long. (accessed 18 Sep 2020).

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122.

Moulson, C. L., Martin, D. R., Lugus, J. J., Schaffer, J. E., Lind, A. C., and Miner, J. H. (2003). Cloning of wrinkle-free, a previously uncharacterized mouse mutation, reveals crucial roles for fatty acid transport protein 4 in skin and hair development. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5274–5279. doi: 10.1073/pnas.0431186100

Ng, P. C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2021). *Biostrings: Efficient Manipulation Of Biological Strings. Bioconductor Version: Release (3.12).* doi: 10.18129/B9.bioc.Biostrings

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Schaffer, J. E. (2002). Fatty acid transport: the roads taken. *Am. J. Physiol. Endocrinol. Metab.* 282, E239–E246.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613.

Tao, J., Koster, M. I., Harrison, W., Moran, J. L., Beier, D. R., Roop, D. R., et al. (2012). A spontaneous fatp4/scl27a4 splice site mutation in a new murine model for congenital ichthyosis. *PLoS One.* 7:e50634. doi: 10.1371/journal.pone.0050634

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43, 11.10.1–11.10.33.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Wang, Q., Pierce-Hoffman, E., Cummings, B. B., Alföldi, J., Francioli, L. C., Gauthier, L. D., et al. (2020). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.* 11:2539.

Wei, L., Liu, L. T., Conroy, J. R., Hu, Q., Conroy, J. M., Morrison, C. D., et al. (2015). MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics.* 16:569. doi: 10.1186/s12864-015-1779-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

2.3. Génotypage par puces à SNP basse densité (60K), haute densité (600K) et imputation *(Résumé de travaux)*

2.3.1. Contexte et objectifs

Une des priorités de la filière « œufs » est d'allonger la carrière de production des poules pondeuses de 70 semaines (durée actuelle des carrières dans le monde) à plus de 90 semaines pour des raisons économiques, éthiques et environnementales. La littérature étant rare à ces âges avancés, deux projets, l'un financé par l'ANR (projet EFFICACE [537]), et l'autre, financé par l'Europe (projet GEroNIMO [538]), ont pour objectif de mieux comprendre le contrôle génétique et épigénétique de différents phénotypes d'intérêt chez la poule pondeuse à âge avancée.

Ce projet intègre différentes données génétiques et épigénétiques de type « omiques », générées à différents niveaux imbriqués du foie (conformation de la chromatine, méthylome, transcriptome régulateur – *e.g.*, ARN non-codants, PCG codant des facteurs de transcription – ou non régulateurs – *e.g.*, PCG codant des enzymes, des protéines de structure…) et disponibles sur des effectifs rarement rencontrés dans la littérature, de par leur coût élevé, soit plus de 400 animaux issus d'une population commerciale de poules pondeuses provenant du sélectionneur NOVOGEN. Le foie a été choisi pour ses différentes fonctions liées à nos phénotypes d'intérêt telles que son rôle clef dans la fabrication du jaune d'œuf, l'homéostasie énergétique et la réaction immunitaire. Notons que d'autres tissus ont également été collectés. Ces données moléculaires originales, et déjà disponibles, seront mises en regard avec des phénotypes d'intérêt, en particulier associés au foie, allant de phénotypes élémentaires (ARN codants des protéines) à des phénotypes plus complexes (lipidome et cellules immunitaires du foie, stéatose hépatique, gras corporel, taux de ponte, qualités des œufs). L'analyse de ces données sera réalisée par différentes approches, notamment des approches génétiques de type GWAS. Le projet qui se déroule sur six ans et qui a débuté en 2021, a pour ambition de fournir des connaissances fondamentales sur la composante génétique de la régulation du transcriptome hépatique et de caractères d'intérêt pour la filière œufs ainsi que des outils pour la sélection des poules pondeuses dans un contexte d'allongement de carrière.

Dans ce contexte, l'ensemble des génotypes disponibles obtenus sur cette population de poules pondeuses, référencée ici sous l'acronyme *Novo*, a été analysé. Ces génotypes ayant

été obtenus à des moments différents avec des puces à SNP différentes (60K *vs.* 600K), l'objectif était de générer des génotypes de qualité avec des coordonnées standardisées sur le nouvel assemblage GRCg7b de la poule. En effet, le sélectionneur privé détenteur de cette population a introduit la sélection génomique dans les années 2010 et a donc génotypé en routine un certain nombre d'individus à chaque génération de sélection avec la puce 600k [411], puce haute densité développée à l'époque par ILLUMINA avec l'assemblage du moment galgal5. À partir de 2017, le sélectionneur a développé sa puce (puce 60K), de moyenne densité et donc de coût plus limité, ceci à partir des SNP da la puce 600K.

2.3.2. Démarches et résultats

Le pedigree détaillé de la population Novo est disponible pour la majorité des individus parmi lesquels 14 341 individus ont été génotypés incluant 9 494 mâles et 4 847 femelles. Parmi ces individus, 1 986 (331 mâles et 1 655 femelles) ont été génotypés avec la puce 600k haute densité alors que 12 355 individus (9 160 mâles et 3 192 femelles) l'ont été avec la puce 60K. La première étape a ainsi consisté à convertir les coordonnées des marqueurs des puces 60k et 600K selon l'assemblage de référence GRCg7b. Pour ce faire, l'outil de *remapping* de NCBI [449]a été utilisé avec les paramètres suivants :

- *Assemblage source : GCF_000002315.4 (Gallus_gallus_5.0)*
- *Assemblage cible : GCF_016699485.2 (bGalGal1.mat.broiler.GRCg7b)*
- *Ratio minimum de bases qui doivent être remappés : 1*
- *Rapport maximal pour la différence entre la longueur de la source et de la cible : 1*
- *Permettre le renvoi à plusieurs sites : TRUE*
- *Fusionner les fragments : FALSE*

En parallèle de cela, pour chaque assemblage, les allèles de référence à chaque position étaient extraits afin de pouvoir les comparer avant et après réalignement. Pour 4 743 marqueurs, une discordance des allèles était observée. Parmi eux, 4 289 (90,4 %) correspondaient à l'allèle complémentaire (*e.g.*, un allèle de référence A devenait allèle de référence T). Dans le doute, l'intégralité des 4 743 marqueurs a été retiré. Pour 359 marqueurs, un réalignement a plusieurs localisations était identifié. Deux cas de figures ont été relevés, soit *i)* le marqueur est réaligné sur des zones proches (quelques bases ou dizaines de bases) et peut correspondre alors à des zones à fortes répétitions, soit *ii)* le marqueur

réaligne sur un chromosome, mais également sur un *scaffold*. Dans tous les cas, au vu du nombre de marqueurs concernés et par mesure de précaution, les 359 marqueurs ont été retirés. Au final, un total de 575 058 marqueurs pour la puce 600K et 54 676 pour la 60k ont été conservés lors de la mise à jour en GRCg7b.

La seconde étape a consisté à réaliser un contrôle qualité, à l'aide de PLINK [539], indépendamment pour chaque puce. Les marqueurs ont été étudiés de manière indépendante entre les chromosomes autosomaux et les chromosomes sexuels. Pour les chromosomes autosomaux, l'ensemble des individus a été conservé et les filtres standards, décrits ci-après, ont été appliqués. Pour le chromosome sexuel Z, seuls les individus mâles ont été conservés. Concernant le W, il s'avérait impossible d'appliquer les filtres usuels car, uniquement chez les femelles, un seul allèle est observable. Le choix a ainsi été fait de conserver tous les marqueurs de ce chromosome. De manière successive, les filtres suivants ont été appliqués :

- Nombre de SNP renseignés par individu supérieur à 80% (*CR SNP > 0,80*)
- Nombre d'individus renseignés par SNP supérieur à 80% (*CR indiv > 0,80*)
- Nombre de SNP renseignés par individu supérieur à 95% (*CR SNP > 0,95)*
- Nombre d'individus renseignés par SNP supérieur à 91% (*CR indiv > 0,91*)
- Fréquence allélique de l'allèle mineure supérieur à 5% (*MAF > 0,05*)
- Valeur de la p-value du test de Hardy-Weinberg supérieure à 0,0001 (*HWE > 0,0001*)

Le choix a été fait d'appliquer les filtres sur les SNP et les individus dans un premier temps avec un seuil de 0,8 et dans un second temps avec un seuil à 0,95 dans le but d'augmenter le nombre de SNP et d'individus conservés, ces deux aspects étant liés. Le choix d'un *CR indiv* à 0,91 dans la deuxième phase apparait comme un compromis entre une valeur de CR élevé et la volonté de conserver un nombre important d'individus. L'impact de ces filtres sur le nombre de SNP et d'individus est disponible dans les tables 1 et 2 suivantes, respectivement pour les puces 600K et 60K.

Table 1 – Impact des différents filtres sur le nombre de SNP et d'individus pour la puce 600K

Puce 600k		
Chromosomes autosomaux		
Filtres	**SNP**	**Individus**
	548384	**1986 (331♂, 1655 ♀)**
CR SNP > 0,80	548298 (-86)	
CR indiv > 0,80		1986 (331♂, 1655 ♀) (-0)
CR SNP > 0,95	542192 (-6106)	
CR indiv > 0,91		1986 (331♂, 1655 ♀) (-0)
MAF > 0,05	301312 (-240880)	
HWE > 0,0001	288225 (-13087)	
Chromosome sexuel (Z)		
Filtres	**SNP**	**Individus**
	26671	**331 (331♂)**
CR SNP > 0,80	26657 (-14)	
CR indiv > 0,80		331 (331♂) (-0)
CR SNP > 0,95	26159 (-498)	
CR indiv > 0,91		331 (331♂) (-0)
MAF > 0,05	10807 (-15352)	
HWE > 0,0001	10531 (-276)	
Chromosome sexuel (W) : 3 SNP		

Table 2 – Impact des différents filtres sur le nombre de SNP et d'individus pour la puce 60K

Puce 60k		
Chromosomes autosomaux		
Filtres	**SNP**	**Individus**
	51439	**12355 (9163♂, 3192 ♀)**
CR SNP > 0,80	50972 (-467)	
CR indiv > 0,80		12355 (9163♂ (-3), 3192 ♀) (-3)
CR SNP > 0,95	50875 (-97)	
CR indiv > 0,91		12352 (9160♂, 3192 ♀) (-0)
MAF > 0,05	43655 (-7220)	
HWE > 0,0001	40593 (-3062)	
Chromosome sexuel (Z)		
Filtres	**SNP**	**Individus**
	3237	**9163 (9163♂)**
CR SNP > 0,80	3176 (-61)	
CR indiv > 0,80		9163 (9163♂) (-3)
CR SNP > 0,95	3163 (-13)	
CR indiv > 0,91		9160 (9160♂) (-0)
MAF > 0,05	2074 (-1089)	
HWE > 0,0001	1825 (-249)	
Chromosome sexuel (W) : 0 SNP		

Au final, pour les chromosomes autosomaux, 12 352 individus et 40 593 marqueurs ont été retenus pour la puce 60k et 1 986 individus et 288 225 marqueurs ont été retenus pour la puce 600k. Les pedigrees étant disponibles, une imputation sur la puce 600K des individus ayant passé le contrôle qualité avec la puce 60K a été réalisée en utilisant FImpute2 [540]. Cette imputation a été faite à la fois pour les chromosomes autosomaux et le chromosome sexuel Z. Les erreurs de filiation avec un taux supérieur à 0,05 ont été identifiées (*param : parentage_test /find_match_cnflt /ert_mm=0.05*) et le remplissage aléatoire des allèles a été désactivé. Pour se faire et afin d'augmenter la précision de l'imputation, les 12 352 individus génotypés avec la puce 60K ont été séparés en deux groupes d'effectif équivalent suivant l'ordre chronologique de génotypage. Ainsi, un premier lot de 5 421 individus a été imputé à partir des 1 986 individus génotypés sous puces 600K. Cet ensemble constitué de 7 407 individus a lui-même été utilisé pour imputer le deuxième lot de 6 931 individus. Au final, les génotypes de 14 338 individus, ayant passé le contrôle qualité, étaient disponibles avec la puce 600K. Les individus ont été redivisés par lot de génération de sélection pour les usages du projet.

2.3.3. Limites et perspectives

Ces données de génotypage, effectués en routine par le sélectionneur NOVOGEN, permettront de nourrir différentes analyses, notamment les analyses GWAS prévues dans le cadre du projet ANR EFFICACE [537] et du projet européen GEroNIMO [538]. Cependant, ces puces étant le résultat d'un réalignement de galgal5 à GRCg7b, les micro-chromosomes ajoutés récemment dans l'assemblage GRCg7b ne sont pas couverts, de même que certaines portions de chromosomes annotées uniquement en GRCg7b. D'autres techniques sont alors envisageables, comme le RNAseq présenté en Résultats §2.1 [529]. Si cette technique se révèle plus couteuse, elle est particulièrement intéressante dans le cadre des analyses eQTL dont le phénotype d'intérêt est l'expression des gènes, ce dernier étant de nos jours principalement mesuré par analyses RNAseq. Ainsi, le RNAseq, résultant d'une étape de séquençage, permet de quantifier les expressions des gènes, mais également d'accéder aux variants génétiques des parties exprimées du génome, qui s'avèrent être nombreux et bien répartis sur le génome, et ont ici l'avantage de couvrir des zones non présentes sur les puces.

Ces analyses sont en cours : à ce jour, les données RNAseq de 700 individus de la population d'intérêt ont été analysées. Cette approche a notamment nécessité une adaptation du *pipeline* de détection de variants par RNAseq « *rnavar* » [541] disponible sous « *nf-core* » [452] faite en collaboration avec Mathieu Charles de l'équipe SIGENAE de l'INRAE. En utilisant les paramètres par défaut (analyse uniquement des exons) du *pipeline nfcore/rnavar* et en appliquant les filtres inspirés des travaux que nous avons effectués sur le RNAseq (FS > 30, QD < 2, CR indiv ≥ 75% et CR indiv ≥ 20% avec au moins 10 *reads*), nous avons observé environ 98 000 variants de type SNP avec génotypage de qualité pour au moins 75% des individus. Des études complémentaires sont maintenant nécessaires pour ré-analyser l'ensemble des 700 données de RNAseq mais cette fois en permettant une recherche de variants dans les exons et introns des gènes, ce qui devrait augmenter d'un facteur deux, voire plus, le nombre de SNP si on se réfère à nos travaux présentés en Résultats §2.1 [529]. La possibilité d'imputer une partie des puces 60K avec les données RNAseq est envisagée.

Pour finir, une autre approche, nommé ddRADseq pour « *Double Digest Restriction-Site Associated DNA* » basée sur la réduction du génome par digestion avec deux enzymes de restriction et séquençage de nouvelle génération permet également d'obtenir des milliers de génotypes fiables et bien répartis sur le génome. Cette méthode a été appliquée à la même population par Mathilde Doublet, une doctorante du laboratoire, en vue de tester son intérêt en évaluation génomique. Ces résultats devraient faire prochainement l'objet d'une publication scientifique à laquelle je suis associé pour avoir contribué à la comparaison des SNP détectés avec du DNAseq 20X et pour la recherche de ceux n'étant observés que par une seule des deux méthodes.

3. Annotation des régions régulatrices du génome : applications à la recherche de gènes causaux dans le cadre des analyses QTL

3.1. L'analyse pilote ChickenGTEx : la détection de régions régulatrices du génome au travers de 28 tissus de populations hétérogènes chez la poule *(Résumé d'article)*

Aparté : Cet article incluant une quantité importante d'analyses, seule une sélection des résultats majeurs est présentée. Ce papier étant un travail collaboratif et donc consensuel, des critiques et des interprétations personnelles, non présentes dans le papier originel, ont été ajoutées. Ces remarques sont indiquées par (NP : Note Personnelle). Notons que les méthodes employées n'ont pas été réintroduites dans ce résumé.

3.1.1. Contexte et objectifs

Comme vu précédemment, la poule est une espèce modèle clé sur plusieurs aspects. D'un point de vue économique, elle est l'une des espèces animales domestiquées fournissant le plus de quantité d'aliments riches en protéines dans le monde entier à travers la production de viande et d'œufs [542]. D'un point de vue de la recherche et au vu de ses caractéristiques phylogénétiques, génétiques et physiologiques, elle est également utilisée comme modèle biologique bien établi dans la recherche fondamentale et appliquée [75, 543], les études de domestication, l'édition du génome, la biologie des systèmes, la virologie, l'immunologie, l'oncologie et l'évolution [544–548]. De plus, la poule arbore un éventail important de variations, principalement entraînées par une forte sélection artificielle et une spécialisation des races, permettant d'étudier l'architecture génétique sous-jacente aux caractères complexes. Fort de ce constat, des lignées divergentes pour certains caractères ont été sélectionnées pour étudier l'impact de la sélection polygénique sur les caractères complexes tels que le poids corporel [549] et le picage des plumes [550] par exemple.

Si la poule est l'une des premières espèces d'animaux d'élevage dont le génome a été séquencé [75], l'effort de recherche la concernant persiste. Ainsi, plusieurs études populationnelles portant sur les variations du génome se sont notamment concentrées sur divers aspects de son évolution, incluant la spéciation et la domestication [551–553], les signatures de sélection [552, 554, 555], le brassage génétique et l'introgression [552, 556, 557], la feralisation (recolonisation d'un milieu sauvage par une espèce domestiquée) [558, 559] et l'adaptation phénotypique [560–562]. De manière concomitante, les analyses de

déséquilibre de liaison et les études GWAS ont identifié des dizaines de milliers de loci associés à de nombreux caractères complexes [417, 546, 563, 564]. La plupart des variants génétiques étant non codants, une annotation et caractérisation systématique de ces variants considérés « régulateurs de l'expression des gènes » apparaissent rapidement essentielles pour tenter de comprendre les voies de régulation sous-jacentes à ces caractères [114, 565, 566]. L'analyse des loci en lien avec la variation des niveaux d'expression des gènes (eQTL) se présente actuellement comme une approche de référence pour mesurer les effets de régulation des variants de séquence sur l'expression génique individuelle [567], à l'instar du projet GTEx humain [389, 454] ainsi que le catalogue eQTL humain porté par l'EBI [568]. Cependant, les précédentes études eQTL chez la poule apparaissent limitées, que ce soit en termes de nombre d'individus, de variants, de gènes, ou encore de types de tissus/cellules [569–572].

Dans ce contexte, le projet *Chicken Genotype-Tissue Expression* (ChickenGTEx), inscrit dans le cadre de l'initiative *Farm animal GTEx* (FarmGTEx) a été initié. L'objectif est de construire un panel de référence des variants régulateurs, notamment à partir de données transcriptomiques, et ce, dans des contextes tissulaires et biologiques distincts (*e.g.*, développement, sexe et exposition environnementale). Dans ce projet donc, 7 015 RNAseq couvrant 52 types de tissus/cellules et 2 869 DNAseq provenant de plus de 100 races/lignées dans le monde entier ont été analysées. Par imputation, environ 1,5M de variants génomiques ont été mis en parallèle de cinq phénotypes transcriptomiques (notés molQTL et incluant eQTL, lncQTL, exQTL, sQTL, 3a'QTL) notamment pour 28 tissus où le nombre d'échantillons était suffisant (de 44 à 741). *(NP)* – Afin de mieux comprendre les limites lors des comparaisons ultérieures, les phénotypes moléculaires sont ici explicités. En réalité, ces observations s'opèrent à plusieurs niveaux :

- Les eQTL et lncQTL se réfèrent à la variation des niveaux d'expression des PCG et des lncRNA respectivement, et sont donc à l'échelle de l'entièreté du gène.
- Les exQTL, sQTL et 3a'QTL se réfèrent à la variation des niveaux d'expression des exons, des sites de splicing et des polyadénylations alternatives en 3'UTR respectivement. Ces observations sont donc à des échelles réduites par rapport au eQTL et lncQTL ainsi moins de reads sont considérés ce qui peut jouer sur la puissance de détection – (NP).

La conservation de ces variants régulateurs au sein des tissus ou selon les contextes a été explorée ainsi que leurs potentiels mécanismes d'action moléculaires sous-jacents. Ils ont par ailleurs été mis en parallèle de 180 caractères complexes dont les loci d'intérêt ont été révélés par GWAS. De plus, la régulation des gènes et les implications phénotypiques identifiées chez la poule ont été comparés avec trois espèces de mammifères, incluant l'humain, le bovin et le porc. Dans l'ensemble, cette étude fournit de nouvelles perspectives sur la hiérarchie des effets régulateurs des variations génétiques sur les transcriptomes de la poule et les phénotypes complexes. De plus, l'atlas des variants régulateurs identifiés constitue une ressource importante notamment pour des études d'amélioration génétique, que ce soit en matière de santé, de production et/ou de résilience. Cette ressource est librement accessible en ligne à l'adresse http://chicken.farmgtex.org.

3.1.2. Résultats

L'étude a été menée selon l'assemblage GRCg6a en utilisant les 16 779 modèles de gène PCG issus de Ensembl v102 [573] auxquels ont été ajoutés les 22 792 lncRNA issus de l'annotation enrichie de Jehl et al., 2020. [124]. *(NP)* – Cet atlas correspond à l'atlas enrichi qui a été développé en 2020 pour l'assemblage galgal5 et qui a été converti en GRCg6a. L'assemblage GRCg7b étant postérieur au début des travaux du ChickenGTEx, le nouvel atlas présenté en Résultats §1.1 n'a pas été utilisé – *(NP).*

Expression et variants

Au travers de l'ensemble des 28 tissus (voir Figure 1D), 23 056 gènes sont considérés comme exprimés (TPM ≥ 0.1), soit 94,7 % des PCG présents. Un total de 1 938 PCG ont été considérés comme spécifiques d'un tissu avec leurs fonctions en cohérence avec la biologie des tissus concernés (voir Figure 1B). De plus, 54,7 % de ces PCG ont pu être associés à au moins un *promoter*/*enhancer* tissue spécifique. En moyenne par tissu, 114 gènes ont pu être identifiés comme DE pour le sexe avec, comme attendu, une surreprésentation dans les chromosomes sexuels associée à une compensation incomplète du dosage des chromosomes sexuels chez la poule. D'un point de vue de la co-expression, 3 538 modules ont pu être identifiés, incluant 10 332 PCG non annotés fonctionnellement par la base de données GO. Ces gènes non

annotés présentent généralement une tissu-spécificité plus importante, une expression plus faible, mais également des liens d'orthologies avec l'humain plus faible. La ressource de 7 015 RNAseq étant riche, une analyse de modélisation de transcrits a été menée. 247 383 transcrits associés à 48 800 loci ont été prédits, dont 184 374 transcrits PCG provenant de 17 215 loci, 13 140 transcrits lncRNA provenant de 3 436 loci et 49 869 autres transcrits non-codants provenant de 34 350 loci. Parmi tous ces transcrits prédits, 90 % n'étaient pas présents dans l'annotation de référence et 4 à 10 % correspondaient à de nouveaux loci génomiques. Au niveau des variants identifiés avec ces données RNAseq, environ 9M de SNP ont été détectés au total. Vu ce faible nombre, 2 869 DNAseq représentant la même diversité populationnelle que les données RNAseq ont été analysés (voir Figure 1C), desquels un ensemble partagé de SNP avoisinant les 1,5M a été observé, permettant alors de les imputer sur les données RNAseq. Enfin, concernant les 52 tissus représentés par 8 668 échantillons, 28 tissus contenant plus de 40 individus ont été retenus pour l'analyse des molQTL (voir Figure 1D). La classification des 7 015 RNAseq sur la base de l'expression des gènes codants des protéines (PCG) dans 28 tissus montre par ailleurs un regroupement par type de tissus (voir Figure 1A)

Caractérisation des molQTL

Comme déjà montré dans d'autres contextes, la puissance statistique de détection des molQTL dépend du nombre d'échantillons (voir Figure 2D). Cette observation est confirmée par l'analyse par sous-échantillonnage croissant de détection d'eQTL dans le foie et le muscle pour lesquels plus de 500 échantillons étaient disponibles. Cependant, la majorité des eQTL à fort effet (aFC ≥ 2) était détectée pour un panel de 200 individus (voir Figure 2F-2G). Afin de vérifier l'impact des modèles dans la détection des molQTL, un modèle linéaire mixte a été appliqué et comparé à la régression linéaire telle que réalisée par TensorQTL [415]. Cette analyse montre une forte corrélation entre les deux méthodes, que ce soit pour la significativité ou l'effet des variants détectés appelés variant *leader*. D'autre part, les échantillons de 15 tissus ayant plus de 100 échantillons ont chacun été subdivisés en deux sous-groupes égaux et la reproductibilité des résultats a été testée. Globalement, un haut taux de reproductibilité de détection (de 0.61 à 0.92) était observé ainsi qu'une forte corrélation entre les effets prédits (voir Figure 2H). Au total, 13 983 (92,9 %) des 15 046 PCG testés (eQTL), 11 685 (74,3 %) des 15 720 lncRNA (lncQTL), 124 423 (76,0 %) des 163 812 exons des PCG (exQTL), 9 669 (61,5 %) des 15 405 loci des PCG présentant des événements d'épissage

alternatif (sQTL) et 8 798 (74,1 %) des 11 880 loci des PCG avec un exon 3'UTR (3a'QTL) étaient significativement régulés par au moins un SNP dans au moins un tissu (voir Figure 2A). Les molQTL étaient globalement enrichis autour des TSS et TES même si, comme attendu, les 3a'QTL et sQTL étaient plus enrichis dans les TES et dans les gènes respectivement (voir Figure 2B). De plus, si entre 40 % et 73 % des molQTL semblaient régulés par plusieurs SNP indépendants, seuls 7 % des 3a'QTL semblait être dans ce cas de figure (voir Figure 2C).

Partage limité des molQTL

Sur l'ensemble des 27 203 gènes testés, 16 097 (59,2 %) présentaient des molQTL significatifs pour au moins deux phénotypes moléculaires. Cependant, les déséquilibres de liaison entre les variants *leader* de ces molQTL ainsi que leurs probabilités de colocalisation étaient faibles en moyenne (0,04 à 0,29), suggérant des mécanismes de régulations indépendants (voir Figure 3A). Les variants *leader* des molQTL étaient significativement enrichis dans certains type de séquences régulatrices (celles-ci étant différentes selon le type de molQTL) à savoir des variants synonymes pour les lncQTL et les exQTL (1,67 et 3,03 fois resp.), des variants 5'UTR pour les eQTL et sQTL (1,82 à 3,64 fois resp.), des variants 3'UTR pour les sQTL et les 3a'QTL (2,29 à 3,77 fois resp.), des transcrits non codants pour les 3a'QTL et les exQTL (1,48 fois à 2,36 fois resp.), des variants d'épissage (surtout accepteurs) pour les sQTL (63,97 fois) et enfin des variants *stop retained* pour les 3a'QTL (5.06 fois) (voir Figure 3C). L'ensemble des molQTL était significativement enrichie en premier lieu dans les états chromatiniens de type *promoter* (moyenne de 3,64 fois), puis dans les états de type *enhancer* (moyenne de 1,98 fois) et les îlots ATAC (moyenne de 1,87 fois) (voir Figure 3D). En revanche, ils étaient significativement appauvris dans les états chromatiniens dits réprimés. Notons que 41 à 73 % des paires eQTL-*eGene* se situent dans les mêmes boucles CTCF qui ont été identifiées dans 22 tissus de poule [565]. Ces résultats pourraient indiquer que les eQTL exercent des effets, par exemple en perturbant les TFBS dans les *enhancer* qui interagissent avec les promoteurs par le biais de boucles 3D de la chromatine.

Conservation entre tissus et races des molQTL

À l'échelle des gènes, un total de 27,4 % des lncQTL et 10,6 % des eQTL n'étaient observés que dans un seul tissu. Pour les molQTL détectés à l'échelle d'un exon ou d'une jonction, 32,1 % des sQTL, 25,8 % des exQTL et 29,6 % des 3a'QTL étaient spécifiques d'un tissu.

(NP) – Cependant, bien que non indiqué dans le papier, il ne faut pas sur-interpréter ces différences de pourcentages. En effet, le fait de détecter plus de QTL dans un seul tissu, par exemple, pour les lncRNA par rapport aux PCG, peut venir en partie d'une plus faible puissance de détection des QTL liée à une plus faible expression – *(NP).*

Les eQTL actifs dans un plus grand nombre de tissus présentaient un enrichissement plus important autour du TSS (voir Figure 4C) et une taille d'effet plus petite (voir Figure 4D). Dans la même idée, les eQTL partagés par les tissus tendaient également à être plus enrichis dans les promoteurs, tandis que les eQTL spécifiques d'un tissu étaient plus enrichis dans les *enhancer*. En général, les tissus ayant des fonctions biologiques similaires ont tendance à être regroupés sur la base de la corrélation de l'effet des molQTL, suggérant des mécanismes de régulations communs associés probablement à leurs fonctions communes (voir Figure 4A et 4B). Notons la détection de 59 eQTL ayant des effets directionnels opposés sur les mêmes gènes (n = 51) entre les tissus. Enfin, la reproductibilité entre races des eQTL a été examinée notamment dans le cerveau, le foie, le muscle et la rate, car tous ces tissus comportaient plus de deux races et chacun d'entre eux avait une taille d'échantillon > 40. La majorité des eQTL (81 % en moyenne) ont pu être reproduits entre les races et le taux de reproduction était d'autant plus grand que l'était la taille de l'échantillon de tissu. En outre, l'effet des eQTL partagés entre races était similaire entre les races (voir Figure 4E).

Dépendance du contexte de molQTL

Pour explorer l'impact des conditions sur la régulation des gènes, les eQTL interagissant avec le sexe (sb-eQTL, pour *sex-biased-eQTL*) et les facteurs de transcription (TF-eQTL) ont été étudiés. Pour la cartographie sb-eQTL, huit tissus ont été considérés où chaque sexe disposait de données provenant de plus de 30 individus. Au total, 1 138 SNP ont révélé une régulation dépendante du sexe pour 962 sb-*eGene*, allant de 3 dans l'intestin grêle à 954 dans le foie. Ces sb-*eGene* détectés dans le sang, l'hypothalamus et le foie étaient significativement enrichis pour les processus biologiques liés au métabolisme des acides aminés, à la voie de transduction des signaux et au métabolisme des acides gras. De même, en examinant 956 TF

chez la poule, une moyenne de 1 941 TF-eQTL dans 17 tissus a été identifiée pour un total de 503 TF. Ces résultats soulignent la dynamique des effets régulateurs génétiques dans des contextes biologiques distincts.

Régulation génétique et caractères complexes

Les molQTL ont été confrontés aux résultats GWAS de 108 caractères complexes, associés à la croissance et le développement (n = 43), la carcasse (n = 41), la production d'œufs (n = 20), l'efficacité alimentaire (n = 3) et l'indice biochimique sanguin (n = 1). L'analyse d'enrichissement a révélé que les loci GWAS de tous les caractères étaient significativement enrichis pour les cinq types de molQTL. Sur les 1 176 loci GWAS significatifs, 1 059 (90 %) pourraient être expliqués par au moins un molQTL dans un des 28 tissus (voir Figure 5A). Sur les 896 loci GWAS colocalisant avec un molQTL, 40 % des gènes les plus proches des variants GWAS correspondaient au gène régulé par le molQTL d'un des 28 tissus (voir Figure 5B).
(NP) – Notons que cette analyse ne prend pas en compte le LD entre le variant *leader* GWAS et les variants de la région qui s'ils sont en très fort LD, pourraient être en réalité le variant d'intérêt – *(NP)*. Ainsi, sur l'ensemble des 1 176 loci GWAS, 0,8 %, 0,9 %, 5,8 % et 1,4 % ont été expliqués uniquement par des eQTL, des sQTL, des exQTL et des lncQTL, respectivement avec toujours la même limitation quant à l'approche. Pour explorer plus avant ces co-localisations molQTL et QTL des caractères complexes, une analyse de co-localisation entre les loci GWAS et les deux types d'eQTL d'interaction avec le contexte détectés en amont a été réalisée. Sur les 1 155 loci GWAS, 22,9 % (264) et 48,7 % (562) ont été expliqués par les sb-eQTL et TF-eQTL respectivement.

Comparaison entre la poule et les mammifères

Sur la base de l'orthologie des gènes entre la poule et trois mammifères (bovins, porcs et humains), il apparaissait que les niveaux d'expression des gènes orthologues (1-1-1-1) étaient significativement plus élevés que ceux des gènes non-orthologues dans l'ensemble des tissus. Après classification sur la base des profils d'expression génique, les 14 278 échantillons tissulaires incluant les quatre espèces ont été regroupés en premier lieu selon le tissu indiquant la conservation au cours de l'évolution de l'expression génique tissulaire (voir Figure 6A). Ce regroupement par tissu a également été confirmé par une forte corrélation des valeurs de tissu-spécificité entre espèces. Comme déjà reportés dans les travaux du PigGTEx [574] et

CattleGTEx [575], la distance entre le variant *leader* du eQTL et le TSS du gène régulé était plus grande chez la poule que chez l'humain. Cela peut s'expliquer en partie par un plus grand LD dans les génomes des animaux domestiqués et à la plus faible densité de SNP des GTEx porc et bovin correspondant à la phase pilote de FarmGTEx par rapport au GTEx humain. Ont également été comparés les 3 024 sTWAS (*single-tissue transcriptome-wide association study*) de 108 caractères chez la poule avec les 9 112, 1 032 et 6 480 sTWAS des trois espèces de mammifères, représentant respectivement 268, 43 et 135 caractères complexes. Pour les tissus correspondants, un total de 8 312 paires de caractères présentant des corrélations significatives entre la poule et une des trois espèces de mammifères a été identifié. La plupart des traits significativement corrélés entre espèces pouvaient s'expliquer par des connaissances biologiques et physiologiques connues. Par exemple, le poids corporel des poules (BW) a montré une forte corrélation avec le gain journalier moyen (ADG) des porcs dans l'iléon et le diabète de type 2 humain (T2D) dans les reins. Ceci était conforme aux résultats précédents selon lesquels d'une part, une plus grande fluctuation de la masse corporelle était liée à un risque accru de T2D chez l'humain [576], et d'autre part, l'expression de ABCC13 dans l'iléon était significativement associée à la fois au BW des poules et à l'ADG des porcs [577]. Les résultats de l'étude TWAS du FarmGTEx offrent ainsi de nouvelles possibilités d'explorer, entre espèces, la conservation de la composante génétique (gènes orthologues) contribuant aux caractères complexes.

3.1.3. Discussions et conclusion

La ressource ChickenGTEx constitue le premier atlas chez la poule de variants régulateurs de l'expression des gènes, de l'épissage alternatif et de la polyadénylation. Cette ressource est librement accessible via un portail web qui inclut des outils de visualisation pour de nombreuses données. Le panel de référence actuel se compose d'environ 3 000 échantillons DNAseq provenant de diverses populations issues du monde entier, offrant ainsi aux chercheurs une ressource de variants et de leurs effets prédits ou encore une ressource permettant d'imputer des génotypes sur 1,5M de variants à partir de ceux dérivés de RNAseq, de puces de SNP ou de séquences à faible couverture. Cette ressource permet de mettre en parallèle de traits complexes d'intérêt, *(NP)* – cependant avec les limites liées à l'imputation de SNP multi-race sur une race précise – *(NP)*.

D'autre part, l'analyse de différents phénotypes moléculaires pour un même gène montre une régulation de ce dernier par des loci génomiques distincts, ce qui est conforme à la complexité déjà connue de la régulation de l'expression génique au niveau *i)* des promoteurs ; *ii)* des *enhancer* impliquant en *trans* des régulateurs de types TF ou lncRNA ou en *cis* des variations génétiques impactant des variations de méthylations ; *iii)* des régions 3'UTR via en particulier l'action en *trans* de miRNA ou encore via l'épissage générant ainsi des isoformes de transcrits différentes. Ce résultat est cohérent avec les résultats obtenus chez l'humain, selon lesquels la plupart des sQTL et des 3a'QTL étaient distincts des eQTL [578, 579]. En outre, un partage élevé de l'effet des eQTL entre les tissus a été observé chez la poule, même s'il est intéressant de noter que c'est le sang qui présente la plus grande dissimilarité par rapport aux autres types de tissus. Cette observation contraste avec celle des mammifères, où les testicules présentent la plus grande dissimilarité [389, 574, 575], ce qui est peut-être dû à la présence de globules rouges nucléés dans le sang aviaire [580, 581].

De plus, un ensemble de variants génétiques régulateurs semblent être en interaction avec les contextes biologiques, par exemple le sexe et l'expression des facteurs de transcription. Ces molQTL dépendant du contexte expliquent 10 à 50 % des loci GWAS, ce qui révèle la nécessité de prendre en compte/standardiser les types/états cellulaires à différents stades de développement, la nutrition et l'état physiologique dans les dispositifs de cartographie molQTL et QTL.

Comme ce fut démontré dans des études humaines [582, 583], l'accumulation des données provenant de diverses races/lignées de poules a augmenté le pouvoir de détection des molQTL en augmentant la taille de l'échantillon, en facilitant la cartographie fine des variants causaux par la réduction du LD des SNP, ainsi qu'en permettant une cartographie des molQTL spécifique à la race [584].

Dans la phase pilote actuelle, les eQTL ayant un effet *trans*(*distant*) (> 1 Mb au TSS des gènes) ne sont pas pris en compte en raison de la taille limitée du dispositif. La découverte de *trans*-eQTL, qui ont souvent une petite taille d'effet, nécessite des centaines de milliers d'échantillons [389, 583], cela sera pris en compte à l'avenir lorsque le nombre d'individus avec données transcriptomiques sera suffisant. En effet, dans le GTEx humain comptant 17 382 individus pour 52 tissus, seuls 143 *trans*-eQTL ont pu être identifiés.

Cette ressource de régulateurs géniques multi-tissus propose une aide à la compréhension génétique et moléculaire des caractères complexes. Environ 90 % des loci GWAS testés dans cette étude sont expliqués par au moins un type de molQTL, une proportion semblable à celle retrouvée chez l'homme (78 %) [389] ou le porc (80 %) [574]. Ce résultat démontre l'usage potentiel de la cartographie molQTL dans la dissection génétique des caractères importants pour les espèces d'élevage, avec un potentiel d'accélération et d'amélioration des programmes actuels de sélection [563, 585]. En effet, au cours des dernières décennies, les études de comparaison entre espèces se sont principalement concentrées sur la séquence d'ADN en raison du manque de données fonctionnelles. Le ChickenGTEx offre donc de nouveaux moyens d'explorer autrement la conservation d'éléments entre espèces en comparant les variants régulateurs sur des caractères complexes et en les traduisant en gènes orthologues comme illustré dans les résultats.

De nouveaux assemblages chez la poule, avec une représentation plus complète, sont maintenant disponibles avec moins de limitations que celles rencontrées avec le génome de référence GRCg6a (Ensembl version 102) [114, 586–588]. De plus, les études futures prendront en compte les *long-read* pour mieux résoudre, entre autres, les variants d'épissage [372, 589, 590]. Il serait également intéressant d'étudier les impacts fonctionnels des variants rares d'une part et des variants somatiques d'autre part sur les phénotypes moléculaires.

Au-delà du transcriptome global, d'autres caractéristiques moléculaires pourraient être incluses, par exemple la variation de la méthylation de l'ADN, la variation des miRNA,

phénotypes impactant l'expression des gènes, ou à un autre niveau, situé au-delà de la transcription, l'abondance des protéines liée au processus de traduction des transcrits, processus qui peut être régulé au niveau génétique (voir discussion générale). Il pourrait également être rajouté d'autres macro-phénotypes tels que les profils métaboliques ou encore, la composition du microbiome qui peuvent interagir plus ou moins directement avec le transcriptome et/ou les caractères d'intérêt. Pour finir, notons qu'il est essentiel d'effectuer des validations expérimentales pour confirmer le statut causal des variants régulateurs candidats et des gènes régulés identifiés dans les signaux GWAS du caractère complexe d'intérêt, ce qui est encore très difficile à grande échelle.

Si cet atlas GTEx de variants régulateurs est associé aux tissus, un atlas sur données *single-cell* pourrait être intéressant à générer pour chaque tissu de façon à explorer de manière plus fine la régulation génique spécifique de populations cellulaires particulières du tissu d'intérêt. En effet, la quantification des proportions des populations cellulaires est rendu possible sur un grand nombre d'échantillons RNAseq via des approches de déconvolution *in silico* à partir d'un petit nombre de données *single-cell* [591].

3.1.4. Valorisation associée

Ces travaux ont fait l'objet :

- d'un article en relecture par les paires : Guan D, Bai Z, Zhu X, Zhong C, Hou Y, Consortium T. C, Lan F, Diao S, Yao Y, Zhao B, Zhu D, Li X, Pan Z, Gao Y, Wang Y, Zou D, Wang R, Xu T, Sun C, Yin H, Teng J, Xu Z, Lin Q, Shi S, Shao D, **Degalez F**, Lagarrigue S, Wang Y, Wang M, Peng M, Rocha D, Charles M, Smith J, Watson K, Buitenhuis A. J, Sahana G, Lund M. S, Warren W, Frantz L, Larson G, Lamont S. J, Si W, Zhao X, Li B, Zhang H, Luo C, Shu D, Qu H, Luo W, Li Z, Nie Q, Zhang X, Zhang Z, Zhang Z, Liu G. E, Cheng H, Yang N, Hu X, Zhou H, Fang L. – The ChickenGTEx Consortium. (2023). The ChickenGTEx pilot analysis: a reference of regulatory variants across 28 chicken tissues. doi: 10.1101/2023.06.27.546670. **Cet article a été soumis à Nature Genetics. En attendant son traitement, il a été déposé sur bioRxiv. Il est reproduit ci-après ;**

Title page

The ChickenGTEx pilot analysis: a reference of regulatory variants across 28 chicken tissues

Dailu Guan[1†], Zhonghao Bai[2†], Xiaoning Zhu[3†], Conghao Zhong[4†], Yali Hou[5,6†], The ChickenGTEx Consortium, Fangren Lan[4], Shuqi Diao[7], Yuelin Yao[8,9], Bingru Zhao[10], Di Zhu[2,3], Xiaochang Li[4], Zhangyuan Pan[11], Yahui Gao[7,12,13], Yuzhe Wang[3], Dong Zou[5], Ruizhen Wang[5,6], Tianyi Xu[5], Congjiao Sun[4], Hongwei Yin[14], Jinyan Teng[7], Zhiting Xu[7], Qing Lin[7], Shourong Shi[15], Dan Shao[15], Fabien Degalez[16], Sandrine Lagarrigue[16], Ying Wang[1], Mingshan Wang[17], Minsheng Peng[17], Dominique Rocha[18], Mathieu Charles[18], Jacqueline Smith[19], Kellie Watson[19], Albert Johannes Buitenhuis[2], Goutam Sahana[2], Mogens Sandø Lund[2], Wesley Warren[20], Laurent Frantz[21,22], Greger Larson[23], Susan J. Lamont[24], Wei Si[11,25], Xin Zhao[25], Bingjie Li[26], Haihan Zhang[27], Chenglong Luo[28], Dingming Shu[28], Hao Qu[28], Wei Luo[28], Zhenhui Li[7,29], Qinghua Nie[7,29], Xiquan Zhang[7,29], Zhe Zhang[7], Zhang Zhang[5,6], George E. Liu[12], Hans Cheng[30], Ning Yang[4]*, Xiaoxiang Hu[3]*, Huaijun Zhou[1]*, Lingzhao Fang[2]*

[1]Department of Animal Science, University of California, Davis, CA, 95616, USA

[2]Center for Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, 8000, Denmark

[3]State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, Beijing, 100193, China

[4]College of Animal Science and Technology, China Agricultural University, Beijing, 100193, China

[5]Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China

[6]University of Chinese Academy of Sciences, Beijing 100049, China

[7] State Key Laboratory of Livestock and Poultry Breeding, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China

[8]MRC Human Genetics Unit at the Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh EH4 2XU, UK

[9]School of Informatics, The University of Edinburgh, Edinburgh EH8 9AB, UK

[10]Jiangsu Livestock Embryo Engineering Laboratory, College of Animal Science and Technology, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China

[11]Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, China

[12]Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, Maryland 20705, USA

[13]Department of Animal and Avian Sciences, University of Maryland, College Park, Maryland 20742, USA

[14]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518124, China

[15]Poultry Institute, Chinese Academy of Agricultural Sciences, Yangzhou, Jiangsu 225125, China

[16]PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France.

[17]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China

[18]Paris-Saclay University, INRAE, AgroParisTech, GABI, Jouy-en-Josas, 78350, France

[19]The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian EH25 9RG, UK

[20]Department of Animal Sciences, Data Science and Informatics Institute, University of Missouri, Columbia, MO 65201

[21]Palaeogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University, Munich 80539, Germany

[22]School of Biological and Behavioural Sciences, Queen Mary University of London, London E1 4DQ, United Kingdom

[23]The Palaeogenomics & Bio-Archaeology Research Network, School of Archaeology, University of Oxford, Oxford, UK

[24]Department of Animal Science, Iowa State University, Ames, Iowa 50011, USA

[25]Department of Animal Science, McGill University, Quebec, H9X 3V9, Canada

[26]Scotland's Rural College (SRUC), Roslin Institute Building, Midlothian EH25 9RG, UK

[27]College of Animal Science and Technology, Hunan Agricultural University, Changsha 410128, China

[28]State Key Laboratory of Swine and Poultry Breeding Industry, Guangdong Key Laboratory of Animal Breeding and Nutrition, Institute of Animal Science, Guangdong Academy of Agricultural Sciences, Guangzhou, 510640, Guangdong China

[29]Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, and Key Laboratory of Chicken Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science, South China Agricultural University, Guangzhou, Guangdong, China

[30]Avian Disease and Oncology Laboratory, USDA, ARS, USNPRC, East Lansing, MI, USA

Corresponding authors*:

Dr. Lingzhao Fang: Center for Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark

E-mail: lingzhao.fang@qgg.au.dk;

Prof. Huaijun Zhou: Department of Animal Science, University of California, Davis, CA, 95616, USA

E-mail: hzhou@ucdavis.edu;

Prof. Xiaoxiang Hu: State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, Beijing, 100193, China

E-mail: huxx@cau.edu.cn;

Prof. Ning Yang: College of Animal Science and Technology, China Agricultural University, Beijing, 100193, China

E-mail: nyang@cau.edu.cn;

Emails for all authors:

dguan@ucdavis.edu; zhonghao.bai@qgg.au.dk; zhu0506@cau.edu.cn; chzhong@cau.edu.cn; houyl@big.ac.cn; fangrlan@163.com; saradiao@126.com; zoud@big.ac.cn; s1914230@ed.ac.uk; zhaobru@163.com; zhudi@qgg.au.dk; lixc1219@126.com; zhypan01@163.com; gyhalvin@gmail.com; yuzhe891@cau.edu.cn; zoud@big.ac.cn;

wangrz@big.ac.cn; xuty@big.ac.cn; cjsun@cau.edu.cn; yinhongwei@caas.cn; kingyan312@live.cn; zhitingxu@126.com; qing_lin1996@126.com; ssr236@163.com; yzshaodan@163.com; fabien.degalez@inrae.fr; sandrine.lagarrigue@agrocampus-ouest.fr; ucywang@ucdavis.edu; wangmingshan@mail.kiz.ac.cn; pengminsheng@mail.kiz.ac.cn; dominique.rocha@inrae.fr; mathieu.charles@inrae.fr; jacqueline.smith@roslin.ed.ac.uk; kellie.watson@ed.ac.uk; bart.buitenhuis@qgg.au.dk; mogens.lund@qgg.au.dk; goutam.sahana@qgg.au.dk; warrenwc@missouri.edu; laurent.frantz@lmu.de; laurent.frantz@palaeo.vetmed.uni-muenchen.de; greger.larson@arch.ox.ac.uk; sjlamont@iastate.edu; siwei01@caas.cn; xin.zhao@mcgill.ca; bingjie.li@sruc.ac.uk; zhhous@163.com; chenglongluo1981@163.com; shudm@263.net; qhw03@163.com; luowei198891@163.com; lizhenhui@scau.edu.cn; nqinghua@scau.edu.cn; xqzhang@scau.edu.cn; zhezhang@scau.edu.cn; zhangzhang@big.ac.cn; george.liu@usda.gov; hcheng@msu.edu; hans.cheng@usda.gov; nyang@cau.edu.cn; huxx@cau.edu.cn; hzhou@ucdavis.edu; lingzhao.fang@qgg.au.dk

Abstract:

Chicken is a valuable model for understanding fundamental biology, vertebrate evolution and diseases, as well as a major source of nutrient-dense and lean-protein-enriched food globally. Although it is the first non-mammalian amniote genome to be sequenced, the chicken genome still lacks a systematic characterization of functional impacts of genetic variants. Here, through integrating 7,015 RNA-Seq and 2,869 whole-genome sequence data, the Chicken Genotype-Tissue Expression (ChickenGTEx) project presents the pilot reference of regulatory variants in 28 chicken tissue transcriptomes, including millions of regulatory effects on primary expression (including protein-coding genes, lncRNA and exon) and post-transcriptional modifications (alternative splicing and 3' untranslated region alternative polyadenylation). We explored the tissue-sharing and context-specificity of these regulatory variants, their underlying molecular mechanisms of action, and their utility in interpreting adaptation and genome-wide associations of 108 chicken complex traits. Finally, we illustrated shared and lineage-specific features of gene regulation between chickens and mammals, and demonstrated how the ChickenGTEx resource can further assist with translating genetic findings across species.

One-Sentence Summary:

The ChickenGTEx provides a multi-tissue reference of regulatory variants for chicken genetics and genomics, functional genomics, precision breeding, veterinary medicine, vertebrate evolution and even human biomedicine.

Main Text:

Introduction

The chicken (*Gallus gallus domesticus*) is not only a globally significant source of protein-rich food through both meat and egg production, but also a fundamental model species. In 2021, the farming industry achieved a staggering production of 111 million tons of eggs and 137 million tons of poultry meat worldwide (https://www.fao.org). Due to its distinct phylogenetic placement as well as its genetic and physiological characteristics, the chicken is also served as a well-recognized model organism in both fundamental and applied research (*1*, *2*), studies of domestication, genome editing, system biology, virology, immunology, oncology, and evolution (*3–7*). The chicken retains a remarkable range of phenotypic variation in terms of morphology, physiology, and behavior, primarily driven by artificial selection and breed specialization. Such extensive variation for a wide range of features is ideal for investigating the genetic architecture underlying complex traits. One example of such traits is dwarfism, which is characterized by a short stature and is observed in various forms in chickens, including sex-linked dwarfism, autosomal dwarfism, and the bantam phenotype, according to their physiological and genetic properties (*8*, *9*). In addition, long-term bidirectional selection lines have been established in chickens to study how polygenetic selection influences complex traits such as body weight (*10*) and feather pecking (*11*).

The Red Jungle Fowl (*G. gallus*), the ancestor of domestic chicken, was one of the first food-producing animals that had its genome assembled (*1*). Recently, a near complete version of the chicken reference genome assembly was reported, revealling distinct sequence and epigenetic features of microchromosomes (*12*). Several population-scale studies of chicken genome variation have focused on various aspects of its evolution, including speciation and domestication (*13–15*), signatures of selection(*14*, *16*, *17*), admixture and introgression(*14*, *18*, *19*), feralization(*20*, *21*), and phenotypic adaptation(*22–24*). Meanwhile, linkage mapping and genome-wide association studies (GWAS) have identified tens of thousands of genomic loci associated with numerous complex traits in chickens (*5*, *25–27*). As most genetic variants behind such adaptive evolutionary and complex traits are non-coding, a systematic annotation of regulatory variants in the chicken genome becomes indispensable for understanding their underlying genetic regulatory circuitry (*28–30*). The expression quantitative trait locus (eQTL) analysis is presently the most powerful approach to measure regulatory effects of sequence variants on individual gene expression in their native genomic and cellular contexts (*31*), as documented in the human Genotype-Tissue Expression (GTEx) project series of studies (*32–34*) and eQTL Catalogue in humans (*35*). In contrast, previous eQTL studies in chickens have been limited in sample size, the number of studied genomic features, and tissue/cell types (*36–40*). For instance, in an intercross population of 125 chickens, Johnsson et al. (2015) identified 6,318 *cis*-eQTL that influence female femoral gene expression, as measured by microarrays (*36*).

To fully unlock the genetic code of the chicken genome, the Chicken Genotype-Tissue Expression (ChickenGTEx) project, as part of the international Farm animal GTEx (FarmGTEx) initiative (*41*), has been lunched to build a comprehensive reference panel of regulatory variants based on the chicken transcriptome in various biological contexts (e.g., development, sex and environmental exposure). In this pilot study, through analyzing 7,015 bulk RNA-Seq datasets from 52 tissues/cell types (hereafter referred to as "tissues") and 2,869 whole genome sequences (WGS) from over 100 breeds/lines (hereafter referred to as "breeds") worldwide, we

systematically associated approximately1.5 million genomic variants with five transcriptomic phenotypes in 28 chicken tissues with sufficient sample size (ranging from 44 to 741). We then explored tissue-sharing and context-dependent patterns of these regulatory variants, their underlying molecular mechanisms of action, and their utility in deciphering GWAS loci of 108 complex traits *via* multiple complementary integrative methods such as transcriptome-wide association studies (TWAS), colocalization, and Mendelian Randomization (MR). Additionally, we compared gene regulation and the phenotypic implications between chicken and three mammalian species (*i.e.* human, cattle and pig). Altogether, our study provides novel and profound insights into the regulatory hierarchy of genetic variation in chicken transcriptomes and complex phenotypes, providing the first large-scale mapping of regulatory variants in the chicken genome and their links to complex phenotypes. Meanwhile, the atlas of regulatory variants identified in this study will facilitate the genetic improvement of chicken populations worldwide in health, production, and resilience and inform a wide range of genetic and genomic research in animal and plant species. Furthermore, we have also well-developed a ChickenGTEx online resource that is freely accessible at http://chicken.farmgtex.org.

Results

Harmonizing large transcriptome and genome datasets in chickens

We analyzed 8,668 bulk RNA-Seq samples using a uniform pipeline, yielding 304.4 billion clean reads. After filtering out low-quality data, 7,015 samples remained for subsequent analyses, representing 28 tissues (**fig. S1, Tables S1** and **2**). Based on gene expressions, samples were clustered well regarding their tissue types (**fig. 1a, fig. S2**). Across all the tissues, an average of 23,056 (94.7% of all annotated genes) genes were expressed (Transcripts per Million, TPM > 0.1) (**Table S3**), showing patterns of ubiquitous or tissue-specific expression (**fig. S3**). An average of 1,938 tissue-specific genes were then detected across tissues (**fig. S3d**), and their functions recapitulated the known tissue biology (**fig. 1b**, **fig. S3e**, **Table S4, URL**). For instance, a total of 1,425 genes were specifically and highly expressed in the bursa of Fabricius, a bird-specific primary lymphoid organ, which were significantly enriched in the immune response to bacteria (**Table S4**). An average of 54.7% of tissue-specific genes could be linked to at least one tissue-specific promoter/enhancer (**fig. S4a-d**). For instance, *MSLNL* with bursa-specific promoters and enhancers showed a specific expression in the bursa (**fig. S4d**). An average of 114 genes exhibited sex-biased expression across 18 tissues (FDR < 0.01), among which 17 genes were shared in all these tissues and located in sex chromosomes (**figs. S4e-f, Table S5**). This was in agreement with the notion of incomplete sex-chromosome dosage compensation in chickens(*42*). In addition, out of 45 genes associated with Mendelian traits in chickens(*43*), 41 showed tissue-specific expression (**fig. S5a**). For instance, *SLCO1B3* is the causal gene of blue eggshell in chickens(*44*), which was specifically and highly expressed in the liver and had liver-specific promoters and enhancers (**fig. S5b**).

To further annotate the function of chicken genes with this extensive transcriptome data, we conducted gene co-expression network and transcript assembly analyses. Based on co-expression analysis, we identified 3,583 co-expression modules containing 25,023 genes, 41.3% (10,332) of which were not functionally annotated in the current Gene Ontology (GO) database (**figs. S6**). In the set of 2,940 unannotated protein-coding genes, 56.3% (1,654) were able to be assigned to co-expression modules. Compared to annotated genes, these unannotated genes exhibited more tissue-specificity, lower gene expression level, and small proportion of chicken-human

orthologous genes (**fig. S6d**). For instance, 8 unannotated genes were co-expressed with 12 annotated genes in the muscle, which were significantly enriched in myeloid cell development and erythrocyte differentiation networks (**fig. S6f**). Through the transcript assembly analysis, we predicted 247,383 transcripts at 48,800 loci, including 184,374 protein-coding transcripts derived from 17,215 loci, 13,140 lncRNA transcripts from 3,436 loci, and 49,869 other noncoding RNA transcripts from 34,350 loci (**fig. S7, URL**). Of all these predicted transcripts, 90% were not annotated in the previous reference and 4-10% were even transcribed from novel genomic loci (**fig. S7d** and **g**). For instance, we observed a new transcript on chromosome 2 was highly and specifically expressed in the testis (**fig. S7h**).

To obtain genotypes of these RNA-Seq samples, we called ~9 million single nucleotide polymorphisms (SNPs) from bulk RNA-Seq data using the GATK best practice pipeline(*45*) (**fig. S8**). To impute missing genotypes, we built a chicken multi-breed genotype imputation reference panel consisting of 2,869 global WGS data sets, which had a similar population composition as the RNA-Seq data (**fig. 1c, Table S6**). Adopting a missing rate of 0.6, the imputation accuracy of 1.5 million SNPs reached 97% (**figs. S8b-h**). The independent datasets from three different chicken breeds confirmed a high concordance rate (> 90%) between imputed genotypes and those directly called from WGS data of the same individuals (**fig. S8g**). After removing duplicated samples based on their genetic relatedness, 28 tissues (each consisting of over 40 individuals) were retained for subsequent molecular quantitative trait loci (molQTL) mapping (**fig. 1d**).

Discovery of molQTL

To comprehensively explore the genetic regulation of the chicken transcriptome, we conducted *cis*-molQTL mapping for five molecular phenotypes, including protein-coding gene expression (eQTL), lncRNA expression (lncQTL), exon expression (exQTL), splicing variation (sQTL), and 3'UTR alternative polyadenylation (3a'QTL), across 28 chicken tissues (**fig. 2a, fig. S2, figs. S9-10**). In total, 13,983 (92.9%) of 15,046 tested protein-coding genes, 11,685 (74.3%) of 15,720 lncRNAs, 124,423 (76.0%) of 163,812 exons, 9,669 (61.5%) of 15,405 loci with alternative splicing events, and 8,798 (74.1%) of 11,880 loci with 3'UTR alternative polyadenylation (3'UTR APA) were significantly (FDR < 0.05) regulated by at least one genetic variant in at least one tissue, and are thus referred to as eGenes, lncGenes, exGenes, sGenes and 3a'Genes, respectively. All the molQTL tended to be enriched around transcription start sites (TSS) and transcription end sites (TES), while 3a'QTL and sQTL showed a higher enrichment in TES and gene body, respectively, compared to other molQTL (**fig. 2b**, **figs. S11a-e**). Furthermore, an average of 73.6% (10,288) of eGenes, 40.5% (3,914) of sGenes, 60.7% (75,527) of exGenes, 58.9% (6,886) of lncGenes, and 7.3% (640) of 3a'Genes were regulated by more than one independent variant (eVariants) across tissues (**fig. 2c, fig. S12**). The further fine-mapping analysis for molQTL with SuSiE (*46*) revealed 2,887, 2,366, 2,053, 12,409 and 1,572 potential causal variants for eGenes, sGenes, lncGenes, exGenes and 3a'Genes, respectively (**URL**).

The statistical power of molQTL mapping depends on the sample size of the tissue, similar to findings in other species(*32*, *47*, *48*) (**fig. 2d-g, figs. S13** and **14**). The down-sampling analysis in the liver and muscle further confirmed the relationship between sample size and eQTL discovery power (**fig. 2g**). Most eQTL with large effect (i.e., fold change of expression, aFC > 2) were detectable when sample size reached around 200 (**fig. 2f** and **g**), and eQTL with larger effect were more enriched around TSS (**fig. S14l**) and had lower minor allele frequency (MAF) (**fig.**

S14). In general, the estimated effect size of eQTL was not correlated with their gene expression levels across tissues (**figs. S15 a** and **b**). Of note, chromosome size was significantly and positively correlated with eGene heritability (**fig. 2e**), eGene discovery (**fig. S15c**), and MAF of lead eVariants (**fig. S15d**). This was only observed in chickens and not in mammals (i.e., pig, cattle and human) (**figs. S15e-g**). Such influences of chromosome size on eQTL effects might be due to differences in evolutionary constraints between microchromosomes and macrochromosomes in chickens(*49*), which was further supported by the observation that phastCons scores of lead eVariants were also negatively correlated with chromosome size (**fig. 2e**).

To validate molQTL identified above, we first applied linear mixed model (LMM), by which we observed that the effect size and significance level of genetic variants estimated by the LMM were highly correlated with those estimated by the linear regression, implemented in tensorQTL (*50*) (**fig. S16**). We then randomly and evenly divided samples from 15 tissues with a sample size of over 100 into two subgroups, and then carried out eQTL mapping separately. A high replication rate, measured by π_1(*51*), was observed between subgroups across tissues, ranging from 0.61 in the hypothalamus to 0.92 in the embryo (**fig. 2h**, **fig. S17**). The effect size of eQTL also exhibited a high Spearman's correlation (an average of 0.77 across tissues) between subgroups (**fig. 2h**). Moreover, we observed that effect sizes derived from the eQTL mapping were positively and significantly correlated (an average of 0.52 across tissues) with those from the allele-specific analysis at the same loci (**fig. 2i, Table S8**). Finally, we trained a deep learning model of regulatory effects based on 310 functional epigenomic profiles in chicken *via* DeepSEA (*50*) (**fig. S15h**, **Table S7**), and observed that regulatory variants predicted by DeepSEA were more significantly enriched in eVariants than the expected (**fig. 2j**). Altogether, these results demonstrated the reliability of molQTL identified in this study.

Limited sharing of regulatory mechanisms underlying five molQTL types

Out of all 27,203 tested genes, 16,097 (59.2%) had significant QTL for at least two molecular phenotypes (**fig. S18a**). The LD of lead variants of any two molQTL types from the same genes was low, ranging from 0.04 (exQTL *vs.* 3a'QTL) to 0.29 (exQTL *vs.* lncQTL) (**fig. 3a**). The colocalization analysis further confirmed the limited sharing of regulatory control among these molecular phenotypes (**fig. 3a**), indicative of their distinct genetic regulatory mechanisms. **Fig. 3b** takes *NLRC5* as an example, four molecular phenotypes of which were controlled by distinct genomic loci, and LD between the respective lead variants was lower than 0.07 (**fig. S18b**). Among these molQTL, eQTL and exQTL exhibited a relatively high colocalization probability (average H4 = 0.72) (**fig. 3a**).

To elucidate molecular mechanisms of action behind these molQTL, we examined sequence ontology and multi-omics data in chickens, including 15 chromatin states predicted from 377 epigenetic data sets in 23 tissues(*30*), and 9,898 topologically associating domains (TADs) detected from high-throughput chromosome conformation capture (Hi-C) in three tissues (i.e., muscle, liver and testis)(*52*). As expected, conditionally independent molQTL were significantly enriched with various regulatory DNA sequences, including synonymous variants (1.67-fold in lncQTL to 3.03-fold in exQTL), 5'UTR variants (1.82-fold in eQTL to 3.64-fold in sQTL), 3'UTR variants (2.29-fold in sQTL to 3.77-fold in 3a'QTL), and non-coding (NC) transcripts (1.48-fold in 3a'QTL to 2.36-fold in exQTL). Of note, all five types of molQTL also exhibited a significant enrichment in missense variants (**fig. 3c**), indicating that a fraction of transcriptional

regulatory variants may also alter protein amino acid residues(*53*). Compared to other molQTL, sQTL exhibited a higher enrichment with splicing variants (63.97-fold in splice acceptor, 3.97-fold in splice donor, and 3.89-fold in splice region), while 3a'QTL were more enriched with stop retained (5.06-fold) and 3'UTR variants (3.77-fold) (**fig. 3c**).

All five types of molQTL showed the highest enrichment in promoter-like states (E1-E5, an average of 3.64-fold), followed by enhancer-like states (E6-E10, an average of 1.98-fold) and ATAC islands (E11, an average of 1.87-fold). In contrast, they were significantly depleted from repressed regions (E12-E14) (**fig. 3d**). Compared to active enhancer (E6), super-enhancer (i.e. a cluster of enhancers in close genomic proximity, exhibiting exceptionally high levels of H3K4ac signals(*30*)), had a lower enrichment for all five molQTL, suggesting that they may be under a stronger purifying selection due to their essential roles in gene regulation and cell identity (**fig. S19a**). Among the five types of molQTL, 3a'QTL had the highest enrichment in enhancer-like states and ATAC islands (**fig. 3d**), supporting their high tissue-specificity. A total of 20% of eQTL, 26% of sQTL, 3.4% of lncQTL, 17.9% of exQTL, and 14.5% of 3a'QTL were supported by regulator-gene pairs that were predicted based on the correlation of signal density of regulators and gene expression (**fig. S19b, Table S9)**. By examining 3D looping of chromatin(*52*), we found 20-60% of molQTL-gene pairs located with the same TAD across tissues (**figs. S19 c** and **d**), with the highest enrichment observed at ~400-600kp away from TSS of target genes after accounting for their distance (**fig. S19e**). As expected, 3a'QTL showed the highest enrichment at ~600-1000kb downstream of their target genes (**fig. S19e**). Likewise, 41-73% of eQTL-eGene pairs located in the same CTCF-loops that were identified from 22 chicken tissues (*30*) (**figs. S19f-h**). These results indicate that the long-distance eQTL exert effects possibly through disrupting TFBS in long-distance enhancers that interact with promoters *via* 3D looping of chromatin (**figs. S19c-h**). As shown in **fig. S19i,** eVariant *rs317368746* regulates expression of *TIMM17B* in the brain only, and it resides in a brain-specific enhancer and is located within the same TAD (346kb upstream) as the TSS of *TIMM17B*. Altogether, these results indicate that regulatory variants exert widespread effects on the transcriptome *via* multiple mechanisms such as changing transcript structure, function, stability, transcription/translation rate and chromatin conformation.

Tissue- and breed-sharing of molQTL

All five types of molQTL were either tissue-specific or ubiquitous, among which 3a'QTL and eQTL exhibited the highest and lowest tissue-specificity, respectively (**fig. 4a** and **b, fig. S20a, fig. 21)**. This was also supported by the meta-tissue analysis (**fig. S22)**. In total, 10.6% of eQTL, 32.1% of sQTL, 27.4% of lncQTL, 25.8% of exQTL, and 29.6% of 3a'QTL were active in one tissue only. Of note, eQTL that were active in more tissues showed a higher enrichment around TSS (**fig. 4c**, **fig. S20b**), a smaller effect size (**fig. 4d**) and a higher MAF (**fig. S20c**). Tissue-shared eQTL (i.e., active in at least two tissues, LFSR < 0.05) also tended to be more enriched for promoter-like states, whereas tissue-specific eQTL were more enriched for enhancer-like states (**fig. S20d**). In general, tissues with similar biological functions (e.g., immune tissues) tended to be clustered together based on eQTL effect correlation (**fig. 4a, fig. S21**), which was similar for the remaining four types of molQTL (**fig. 4b, fig. S21)**. Unlike GTEx in mammals(*32*, *47*, *48*), blood formed the primary outgroup in chickens regarding eQTL and lncQTL, while, for the remaining three types of molQTL, brain and testis were first separated from the rest of the tissues. **Fig. S20e** demonstrates an eQTL (9_16035177_G_A) that significantly regulated the expression of *ALG3* only in the blood. The *ALG3* gene encodes alpha-1,3-mannosyltransferase

with the function of inducing glycosylation of TGF-β receptor II(*54*), which might modulate blood pressure homeostasis(*55*) and affect hematopoiesis(*56*). In contrast to eQTL shared in other tissues, blood-specific eQTL had a lower MAF (**fig. S20f**) and a larger effect (**fig. S20g**). Moreover, genetic regulation of all five molecular phenotypes in the embryo was distinct from those in the primary tissues (**fig. 4a, fig. S21**), similar to that in pigs(*47*), indicating a distinct regulation of early development. In addition, we detected 59 eQTL with opposite directional effects on the same genes (n = 51) between tissues (**Table S10**). For instance, the T-allele of *rs315639985* increased the expression of *FBXO5* in the spleen but decreased its expression in the whole blood. The *FBXO5* gene encodes F-box protein 5, which is associated with systolic blood pressure in human(*57*) (**fig. S20h**). Another example was *rs313608694,* whose G-allele significantly upregulated the expression of *ELAC2* in the embryo but downregulated it in the spleen (**fig. S20h**). This gene encodes elaC ribonuclease Z 2, and the reduction of its expression could induce growth arrest by suppressing transforming growth factor-beta(*58*).

We examined breed-sharing of eQTL in the brain, liver, muscle and spleen, as all of them had more than two breeds and each with a sample size > 40. As a result, the majority of eQTL (an average of 81%) could be replicated between breeds and the replication rate was associated with tissue sample size (**fig. S23a, Table S11**). Furthermore, the eQTL effect was substantially shared between breeds (**fig. 4e**). For instance, the T-allele of *rs314795649* significantly upregulated expression of *PRKCDBP* in the liver across all four breeds being tested, including Cobb ($\beta = 0.57$, $P = 2.67 \times 10^{-6}$), Leghorn ($\beta = 0.33$, $P = 3.10 \times 10^{-6}$), Rhode Island Red ($\beta = 0.39$, $P = 3.06 \times 10^{-6}$) and Ross ($\beta = 0.37$, $P = 5.02 \times 10^{-10}$) (**fig. S23b**). In addition, we detected 376 (Red Jungle Fowl vs. Ross) and 185 (Red Jungle Fowl vs. Leghorn) breed-interaction eQTL (bi-eQTL) in the brain, and with genes regulated by them were enriched in functionals related to brain development (**Table S12**).

Context-dependence of molQTL

To explore the context-dependent nature of gene regulation, we systematically detected eQTL interacting with sex (sb-eQTL), transcription factor (TF-eQTL) and cell type (ci-eQTL). For sb-eQTL mapping, we only considered eight tissues, where each sex had data from over 30 individuals available. In total, 1,138 SNPs displayed sex-biased regulation of 962 eGenes (sb-eGene, FDR < 0.01), ranging from 3 in the small intestine to 954 in the liver (**URL**). Taking the liver as an example, we further performed the sb-eQTL mapping in a single breed, Rhode Island Red (n_{male}= 32; n_{female}= 46), resulting in 48 sb-eQTL regulating 30 eGenes (**fig. 4f, figs. S23c-d, Table S13**). For instance, the significant association of *rs317663121* with *TCFL5* expression was only observed in male liver (**fig. 4f**). Moreover, 14% (164) of sb-eGenes overlapped with sex-biased expressed genes in all eight studied tissues. These sb-eGenes detected in the blood, hypothalamus, and liver were significantly enriched in biological processes related to amino acid metabolism, signaling transduction pathway, and fatty acid metabolism (**Table S14**). Through the examination of 956 chicken transcription factors retrieved from the AnimalTFDB 3.0(*59*), we detected an average of 1,941 TF-eQTL in 17 tissues, representing 503 TFs (**fig. S23f, URL**). **Fig. S23e** illustrates that effect of *rs313600592* on *ATP6V1A* expression was significantly associated with the expression of transcription factor *TCF25* in the muscle. For ci-eQTL mapping, we first annotated 13 cell types from single-cell RNA-Seq data in chicken heart and muscle (**Table S15**). Based on the cellular composition of bulk RNA-Seq samples of muscle and heart estimated by the *in silico* cell-type deconvolution (**fig. S24**), we identified an average of 105 ci-eGenes in the muscle, ranging from 11 with interactions in adipocytes to 214 with Schwann cells, and an

393 average of 19 ci-eGenes in the heart, ranging from 6 interacting with fibroblasts to 36 with
394 cardiomyocytes (**fig. S23g, Table S16**). For instance, *rs733070738* regulated expression of
395 *PLVAP* by interacting with myocyte enrichment in the muscle (**fig. 4g**). These results highlight
396 the dynamics of genetic regulatory effects across distinct biological contexts.

397

398 **Interpreting genetic regulation behind complex traits and adaptive evolution**

399 To show the potential of molQTL in understanding complex traits in chickens, we systematically
400 integrated molQTL with GWAS results of 108 complex traits, including
401 growth and development (n = 43), carcass (n = 41), egg production (n = 20), feed efficiency (n =
402 3), and blood biochemical index (n = 1) (**Table S17**). Enrichment analysis revealed that GWAS
403 loci of all the traits were significantly enriched in all five types of molQTL (**fig. S25a**). Among
404 them, the highest enrichment was observed for 3a'QTL (1.87±0.33), followed by sQTL
405 (1.83±0.28), eQTL (1.81±0.30), lncQTL (1.59±0.27) and finally exQTL (1.56±0.32) (**fig. S25a**).
406 Furthermore, we applied four complementary methods to prioritize causal variants and genes
407 underlying each GWAS loci, including fastENLOC-based colocalization, summary-data-based
408 MR (SMR), single-tissue transcriptome-wide association study (sTWAS), and multi-tissue
409 TWAS (mTWAS). Out of all 1,176 significant GWAS loci, 1,059 (90%) could be explained by
410 at least one molQTL across 28 tissues (**fig. 5a, figs. S26** and **S27**). Of 896 colocalized GWAS
411 loci, 59.9% were not colocalized with the nearest genes of lead GWAS variants, indicative of the
412 regulatory complexity of complex traits (**fig. 5b, fig. S28a**). The number of colocalization events
413 of a trait was determined by the statistical power of both GWAS and molQTL mapping (**fig.**
414 **S26b-c**). Of all 1,176 GWAS loci, 0.8%, 0.9%, 5.2% and 1.4% were explained uniquely by
415 eQTL, sQTL, exQTL and lncQTL, respectively. This result indicates that each type of molecular
416 phenotype only had a limited contribution to complex traits at distinct levels of gene regulation
417 (**fig 5a**, **fig. S29**). Taking the body weight gain from week 6 to 8 (WG6.8) as an example,
418 sTWAS linked GWAS loci to 43 unique genes (34 protein-coding and 9 lncRNA genes) across
419 21 tissues (**fig. 5c, fig. S28c, Table S18**). Of them, the expression of the *KPNA3* (karyopherin
420 subunit alpha 3) exhibited the strongest association with WG6.8 in the retina, followed by
421 pituitary and heart (**fig. 5c**). Consistently, it has been documented that the knockdown of the
422 *KPNA3* would restore photoreceptor formation in *Drosophila*(*60*). The highest colocalization
423 between WG6.8 GWAS loci and molQTL of *KPNA3* was observed for a retina eQTL
424 (*rs314814283*, GRCP=0.78) and a pituitary sQTL (*rs13552958*, GRCP=0.54) (**Table S19**). The
425 further SMR analysis pinpointed 10 potential causal mutations across tissues (**Table S20**),
426 among which *rs739579746* was the most significant one (**fig. 5c, Table S20**). The SNPs
427 *rs314814283* and *rs739579746* detected by eQTL mapping were in high LD (r^2 = 0.88), while
428 both showed low LD with *rs13552958* (r^2 < 0.02) detected by sQTL mapping. These findings
429 likely reflect the importance of photoreception for chicken growth and production performance
430 (*61*, *62*), and the promising candidate gene in this region is the *KPNA3*. In addition, we detected
431 149 significant lncRNA-protein-coding-trait regulation events with SMR-multi analysis (**Table**
432 **S21**). For instance, an eQTL of a lncRNA (*ENSGALG00000053557*), located on the opposite
433 strand of the *IL20RA*, exhibited significant colocalizations with an eQTL of *IL20RA* in the
434 muscle and GWAS loci of the total stomach weight on chromosome 3 (**fig. 5d**).

435 To further explore context-specific genetic regulation of complex traits, we conducted
436 colocalization analysis between GWAS loci and three types of context-interaction eQTL
437 detected above (**fig. S25b**). Out of 1,155 GWAS loci, 22.9% (264), 48.7% (562) and 12.3%
438 (142) were explained by sb-eQTL, TF-eQTL and ci-eQTL, respectively (**fig. S25b**). For

instance, GWAS loci of total stomach weight and body weight at 8 weeks of age were significantly colocalized with sb-eQTL of *MFSD4A* and *TOX3* in the brain and spleen, respectively (**figs. S28d-e**). The *TOX3* gene encodes TOX high mobility group box family member 3, playing roles in sex determination and differentiation (*63*, *64*). Despite the limited discovery power of the context-interaction eQTL due to the small sample size, our analysis demonstrated that context-specific regulatory effects were nonnegligible in dissecting the regulatory mechanism of complex traits. Furthermore, we conducted an exploratory analysis to investigate whether domestication and breeding also tend to target on regulatory variants, though examining selection sweeps previously detected between broilers and layers previously (**fig. S25d, fig. S30**) (*14*, *65*). Within the brain, we separately detected eQTL in three chicken lines/breeds separately, including Red Jungle Fowl (n = 46), Ross (n=157) and Leghorn (n = 78). Genomic windows containing at least one eQTL (i.e., eQTL windows) in Ross and Leghorn were under stronger selection (i.e., larger selection values, LSBL) in broilers than expected, whereas those detected in Red Jungle Fowl were not (**fig. S25d, fig. S30**). Likewise, for selection sweeps in layers, eQTL windows in Leghorn were under stronger selection in layers than expected, but not for eQTL windows in Ross and Red Jungle Fowl (**fig. S25d, fig. S30**). Altogether, the current ChickenGTEx can serve as a valuable resource for exploiting regulatory mechanisms underlying complex traits and adaptation in chickens.

Comparing gene regulation and complex trait genetics between chickens and mammals

Based on gene orthology between chickens and three mammals (i.e., cattle, pigs and humans), we found the expression levels of the 1-1-1-1 orthologous genes were significantly higher than those of non-orthologous genes across tissues (**fig. S31**). The proportion of orthologous genes expressed in chicken tissues was positively (Pearson's $r > 0.8$, $P < 0.004$) correlated with that in mammalian tissues (**fig. S31c**). Based on gene expression profiles, 14,278 samples in the four species were clustered first according to their tissue types, indicating the global conservation of gene expression between chickens and mammals (**fig. 6a**). This was also supported by a high correlation of TAU values of genes, a measure of tissue-specificity of gene expression, between chickens and mammals (**fig. S31d**). The phylogenetic analysis of gene expression revealed different evolutionary rates of tissues across species, where testis and pituitary evolved fastest, while adipose and liver evolved slowest (**figs. S31e**). The effect sizes of lead eQTL of orthologous genes were significantly but weakly correlated between chickens and mammals, which were lower than those within mammals (**fig. 6b**). This was consistent for *cis*-h^2 of orthologous genes (**fig. S32a**). As in pigs and cattle, the distance of lead eQTL to TSS was larger in chickens than that in humans (**fig. S32b** and **c**), which might be partially due to the larger LD of SNPs in farm animals' genomes and lower SNP density in the pilot phase of FarmGTEx compared to human GTEx(*32*). We further divided chicken eGenes of each tissue into two groups: 1) chicken-specific eGenes, and 2) those shared with at least one mammalian species (conserved eGenes) (**see Methods**). In general, compared to chicken-specific eGenes, conserved eGenes showed a higher gene expression level, lower tissue-specificity, were more likely to be differentially expressed between species, have more promoters, and stronger tolerance to loss-of-function mutations (less evolutionarily constrained) (**fig 6c**).

The FarmGTEx-based TWAS results provide new opportunities to systematically explore between-species similarity of complex trait genetics at the functional level of orthologous genes. We thus compared all the 3,024 sTWAS of 108 traits in chickens with 9,112, 1,032 and 6,480 sTWAS in three mammalian species, representing 268, 43 and 135 complex traits, respectively.

Within the matching tissues, we identified a total of 8,312 trait-pairs with significant correlations between chickens and three mammalian species ($P < 9.11 \times 10^{-3}$, permutation-based) (**fig. 6d**, **fig. S32d-f, Table S22**), despite the big differences in the TWAS power between species. Most of the significantly correlated traits between species recapitulated known biological and physiological knowledge. For instance, chicken body weight (BW) showed a high correlation with pig average daily gain (ADG) in the ileum (Pearson's $r = 0.69$, $P = 3.42 \times 10^{-5}$, **fig. 32g**), cattle somatic cell scores (SCS) in the adipose (Pearson's $r = 0.38$, $P = 5.46 \times 10^{-5}$, **fig. 32h**), and human type 2 diabetes (T2D) in the kidney (Pearson's $r = 0.57$, $P = 2.3 \times 10^{-5}$, **fig. 32i**). This was in line with previous findings that larger BW fluctuation was related to an increased T2D risk in human(*66*), and also was positively associated with SCS in cattle (*67*). The expression of *ABCC13* (encoding ATP binding cassette subfamily C member 13) in the ileum was significantly associated with both chicken BW ($P = 0.03$) and pig ADG ($P = 0.04$), which encodes ATP binding cassette subfamily C member 13, which had potential associations with body weight/body mass index in humans (*68*). The expression of three genes, *PIGX*, *MRPL51* and *ABHD14B*, in the adipose were significantly associated with both chicken BW and cattle SCS. Of these, the ABHD14B protein is a lysine deacetylase with the capacity of catalyzing the deacetylation of lysine residues to yield acetyl-CoA, which could significantly alter glucose metabolism and could thus cause significant BW loss(*69*, *70*). The expression of *GABRB2* and *SOX4* in the kidney was significantly associated with both chicken BW and human T2D. The SOX4 is involved in pancreas development with roles in inhibiting insulin secretion and increasing diabetes risk(*71*, *72*). Moreover, taking chicken BW as an example, we carried out cross-species meta-TWAS analysis in the muscle, and found that homologous traits (e.g., ADG and back fat thickness) rather than non-homologous traits (e.g., number of stillborn and weaned pigs) in pigs could help detect more genes associated with BW in chickens **(fig. 6e)**. Similarly, human height and BMI increased the detection power of BW-associated genes in chickens *via* cross-species meta-TWAS analysis in the muscle (**fig. S32k**). These results highlighted that the FarmGTEx resource could facilitate the translation of genetic findings between species at the functional level of orthologous genes rather than the DNA sequence level.

Discussions

Summary and general impacts: Through the comprehensive analyses of the so-far largest collection of chicken RNA-Seq and WGS data, we have developed a catalogue of genetic variants with regulatory effects on five transcriptional phenotypes, representing both primary expression (including protein-coding, lncRNA and exon) and post-transcriptional modifications (alternative splicing and 3'UTR APA), across 28 chicken tissues, referred as the ChickenGTEx. We made the findings and resources of ChickenGTEx freely accessible to the entire community through http://chicken.farmgtex.org. This web portal provides an open-access chicken genotype imputation reference panel, which was built-up and maintained as part of this project. The current reference panel consists of approximately 3,000 WGS samples from around the globe, enabling researchers to impute genotypes derived from RNA-Seq, SNP array or low-coverage sequences to the whole-genome sequence level, which can be utilized further to prioritize potential causal variants underlying complex traits of interest through integrating with muti-layer ChickenGTEx resources. Besides, we offer highly-useful visualization tool, Integrative Genomics Viewer (IGV) (*73*) for exploring molecular phenotypes, enhancer-gene interactions, chromatin states, epigenetic modifications, and publicly available GWAS results. The web portal also includes single-cell RNA-Seq data that were collected and analyzed from six chicken

tissues, enabling users to query the expression of their desired genes at both the cellular and bulk tissue level. Additionally, we provide batch data download and advanced search options for data resource generated in this study, and will continue updating the database to ensure its future accuracy and relevance. Overall, this first GTEx resource in avian species serves as a valuable resource for a global atlas of regulatory variants in chickens and informs vertebrate genome evolution at the functional level, benefiting future research in animal, plant, and human genetic and biomedicine research.

MolQTL mapping and the underlying molecular mechanism: We have demonstrated that different molecular phenotypes of the same genes were likely to be controlled by distinct genomic loci through distinct regulatory mechanisms, indicating the importance of integrating omics data corresponding to multi-layer biologically-important molecular phenotypes (e.g., epigenetic mark activity and microRNA expression(*74*)) in future studies. This is consistent with findings in humans that most of the sQTL and 3a'QTL were distinct from eQTL (*75*, *76*). The comparative analysis of regulatory variants reveals several specificities of gene regulation in chickens compared to mammals. For instance, the chicken genome exhibits a chromosome-size dependence in genetic control of gene expression, in contrast to mammals. Avian genomes often have chromosomes of highly variable sizes, with chicken chromosomes ranging from a minimum of 3.4 Mb to a maximum of 200 Mb (*12*). Chicken microchromosomes exhibit a higher gene density, higher GC content and DNA methylation levels(*12*, *77*), and are under stronger evolutionary constraints (*49*), that easily distinguish them from the mammal 'like' macrochromosomes. These distinct genetic and epigenetic features might lead to differences in the genetic regulation of gene expressions across chromosomes in chickens. In addition, we observed a high sharing of eQTL effect across tissues in chickens, while interestingly the blood showed the highest dissimilarity against other tissue types. This observation is in contrast to that of mammals, where the testis showed the highest dissimilarity (*32*, *47*, *48*) that is perhaps a result of nucleated red blood cells in avian blood (*78*, *79*). Moreover, we uncovered a set of genetic variants with regulatory effects interacting with biological contexts, e.g., sex, transcription factor expression, genetic background, and cell type compositions. This context-dependent molQTL explained 10-50% of GWAS loci, revealing the need to consider cell types/states under different developmental stages, nutrition, and physiology status in the future molQTL mapping experiments. By taking account of a wide range of environmental/biological contexts, we can effectively tackle the challenge of "missing regulation" (*80*, *81*). As demonstrated in human studies(*82*, *83*), harmonizing data from diverse chicken breeds/lines increased the detection power of molQTL *via* increasing sample size, facilitating the fine-mapping of causal variants *via* reducing LD of SNPs, as well as allowing breed-specific molQTL mapping (*84*). At the current pilot phase, eQTL with *trans*-regulatory effect (> 1Mb to the TSS of genes) is not considered due to the limited sample size. Discovering *trans*-eQTL, which often has a small effect size, requires hundreds of thousands of samples (*82*, *85*), and will be considered in the future when the sample size of transcriptome data is sufficient.

Potential applications of ChickenGTEx: This multi-tissue gene regulation resource opens the door to decipher the biological mechanism of complex traits, domestication and polygenic adaptation in chickens in-depth. It enables nearly 90% of GWAS loci being tested in this study to be explained by at least of one type of molQTL, a higher proportion than that in humans (78%) (*32*) or in pigs (80%) (*47*). This finding demonstrates the importance of molQTL mapping in functionally dissecting agriculturally important traits in farm animals, with a high potential for accelerating and improving the current animal breeding program and enabling the future precision selection and breeding (*26*, *86*). The focus of cross-species comparison studies in the

past decades was mainly on the DNA sequence level due to the lack of relevant functional data, and the recent Zoonomia project investigated the DNA sequence evolution of regulatory elements while based on *in silico* prediction across species (*87–89*). The ChickenGTEx offers new means to explore the evolutionary impacts of gene regulation on complex traits across species and translate genetic findings between species at the functional level of orthologous genes rather than the DNA sequence level. Our exploratory comparative analysis of large-scale TWAS between chickens and mammals illustrates how to "borrow" information between species for gene mapping (*90*, *91*). We found that cross-species meta-TWAS aided in the identification of more functional genes for homologous traits. We believe that the ChickenGTEx resource will not only contribute significantly to elucidating the molecular architecture underlying phenotypic variation in chickens, but also to developing chicken models for studying human complex traits (e.g., disease and behavior (*3–7*)).

Limitations and outlooks: The current ChickenGTEx provides the most expansive source of regulatory variants in the chicken genome. Some limitations and challenges remain in the genotype and molecular phenotype assessments. New chicken assemblies with more complete representation are becoming available with fewer computational limitations that we experienced using the GRCg6a reference genome (Ensembl version 102) (*12*, *92–94*). Future studies will consider long-read sequences to better resolve splice-variants (*95–97*), and pangenome references to annotate complex structural variants (*98*), mobile element variation (*99*), and short tandem repeats (*92*, *100*). In addition, it would be of great interest to investigate the functional impacts of rare and somatic variants on molecular phenotypes, where multi-tissue samples are collected from the same individuals with deep WGS data available. Beyond the bulk transcriptome, other molecular features could be included, e.g., DNA methylation variation, protein abundance, metabolite profiles, and the composition of the microbiome. For future single-cell genetics in chickens, a comprehensive chicken single-cell atlas will be the first step and is urgently required to explore the cell-type/state-specific gene regulation *via in silico* cell type deconvolution of large bulk tissue samples(*101*). In addition, conducting experimental follow-ups *via* methods, e.g., massively parallel CRISPR-based screens (*102*), is crucial to functionally validate and characterize regulatory effects of genetic variants and to identify functional genes of complex traits on a large-scale. In summary, the current and future versions of the ChickenGTEx project promises to establish a reference panel for studying the functional impacts of genetic variants in their native genomic and cellular contexts in distinct biological contexts, including molQTL mapping, molecular phenotype prediction for individuals with genotypes (including extinct species with ancient DNA available) and the evolution of regulatory variants. The fully developed ChickenGTEx will contribute substantially to research in complex-trait genetics, animal breeding, functional biology, and vertebrate genome evolution at the functional level.

Acknowledgments:

We thank all the researchers who have contributed to the publicly available data used in this research.

Funding:

D. Guan was supported from Agriculture and Food Research Initiative Competitive grants nos. 2020-67015-31175, and 2022-67015-36215 (H.Z.) from the USDA National Institute of Food and Agriculture. H.Z. acknowledges fundings from Agriculture and Food Research Initiative

Competitive grants nos. 2020-67015-31175, 2015-67015-22940, and 2022-67015-36215 (H.Z.) from the USDA National Institute of Food and Agriculture, Multistate Research Project NRSP8 and NC1170 (H.Z.), and the California Agricultural Experimental Station (H.Z.). N.Y. acknowledges fundings from the National Key Research and Development Program of China (2021YFD1300600 and 2022YFF1000204). X. H. was supported by the National Natural Science Foundation of China, "Genetic architecture, gene interaction and genomic prediction for chicken growth evaluated using large advanced intercross populations" (31961133003). Y.W. was supported by the National Natural Science Foundation of China, "Deciphering the genetic architecture of polygenic clustered QTL for chicken body weight by integrative omics", (32272862). S Rong acknowledges funding from Jiangsu Agricultural Industry Technology System (JATS[2022]406). Zhe Zhang acknowledges fundings from the National Natural Science Foundation of China (32022078 to Z.Z.), the Local Innovative and Research Teams Project of Guangdong Province (2019BT02N630 to Q.N. and Z.Z.). Zhang Zhang acknowledges fundings from National Natural Science Foundation of China (32030021), National Key Research & Development Program of China (2021YFF0703702) and Technical Support Talent Program of Chinese Academy of Sciences (awarded to DZ). Y.H. acknowledges funding from the Science and Technology Innovation 2030 - Major Project (2022ZD04017). L.W. H.Q. and C.L., were supported by Science and Technology Planning Project of Guangzhou City (201504010017) and Natural Scientific Foundation of China (31402067). G.E.L. was supported in part by USDA NIFA AFRI grant numbers 2019-67015-29321 and 2021-67015-33409 and the appropriated project 8042-31000-112-00-D, "Accelerating Genetic Improvement of Ruminants Through Enhanced Genome Assembly, Annotation, and Selection" of the USDA Agricultural Research Service (ARS).

Author contribution statement:

L.Fang, H. Zhou, D.G., X.H., and N.Y. conceived and designed the project. D.G., Y.Y., B.Z. and Z.P. performed bioinformatic analyses of RNA-Seq data analysis. D.G., F.L., S.D., Y.G. and H.Y. conducted whole-genome sequence data analysis. D.Z., performed the deep learning analysis. D.G. performed multi-omics and single-cell RNA-Seq data analysis. D.G. conducted molQTL mapping. X.Z., C.Z. D.G. performed GWAS integrative analysis. Z.B. and D.G. led the comparison of GTEx between chickens and mammals. L.F., H.Zhou, D.G., X.Z., Q.L., C.Z., Y.H., Y.W., C.S., J.T., F.D., S.L., Y.W., M.W., M.P., D.R., M.C., J.S., K.W., A.J.B., W.W., L.Frantz, G.L., M.S.L., G.S., S.S., D.S., S.J.L., X.Z., B.L., H.Zhang, and H.C. contributed to the critical interpretation of analytical results before and during manuscript preparation. Y.H., D.Z., R.W., T.X., and Zhang Zhang built the ChickenGTEx web portal. H. Zhou, L.Fang, N.Y., X.H., G.E.L., Zhe Zhang, S.S., D.S., X.Z., Q.N., Z.L., W.L., H.Q., W. S. and C.L. contributed to the data and computational resources. D.G., Z.B., X.Z., C.Z., Y.W., Y.H. and L.Fang drafted the manuscript. All authors read, edited, and approved the final manuscript.

Competing interests:

The authors declare no competing interests.

Data and materials availability:

664 All raw data analyzed in this study are publicly available for download without restrictions from
665 SRA (https://www.ncbi.nlm.nih.gov/sra/) and NGDC BioProject
666 (https://bigd.big.ac.cn/bioproject/) databases. Details of RNA-Seq, WGS, ChIP-Seq peaks and
667 single-cell RNA-Seq can be found in Table S1, S6, S7 and S15, respectively. All processed data
668 and the full summary statistics of molQTL mapping and genotype imputation reference panel are
669 available at http://chicken.farmgtex.org. All the computational scripts and codes for RNA-Seq,
670 WGS, single-cell RNA-Seq and Hi-C datasets analyses, as well as the respective quality control,
671 molecular phenotype normalization, genotype imputation, molQTL mapping, functional
672 enrichment, colocalization, SMR and TWAS are available at the FarmGTEx GitHub website
673 (https://github.com/FarmOmics/ChickenGTEx_pilot_phase).

Figures and legends

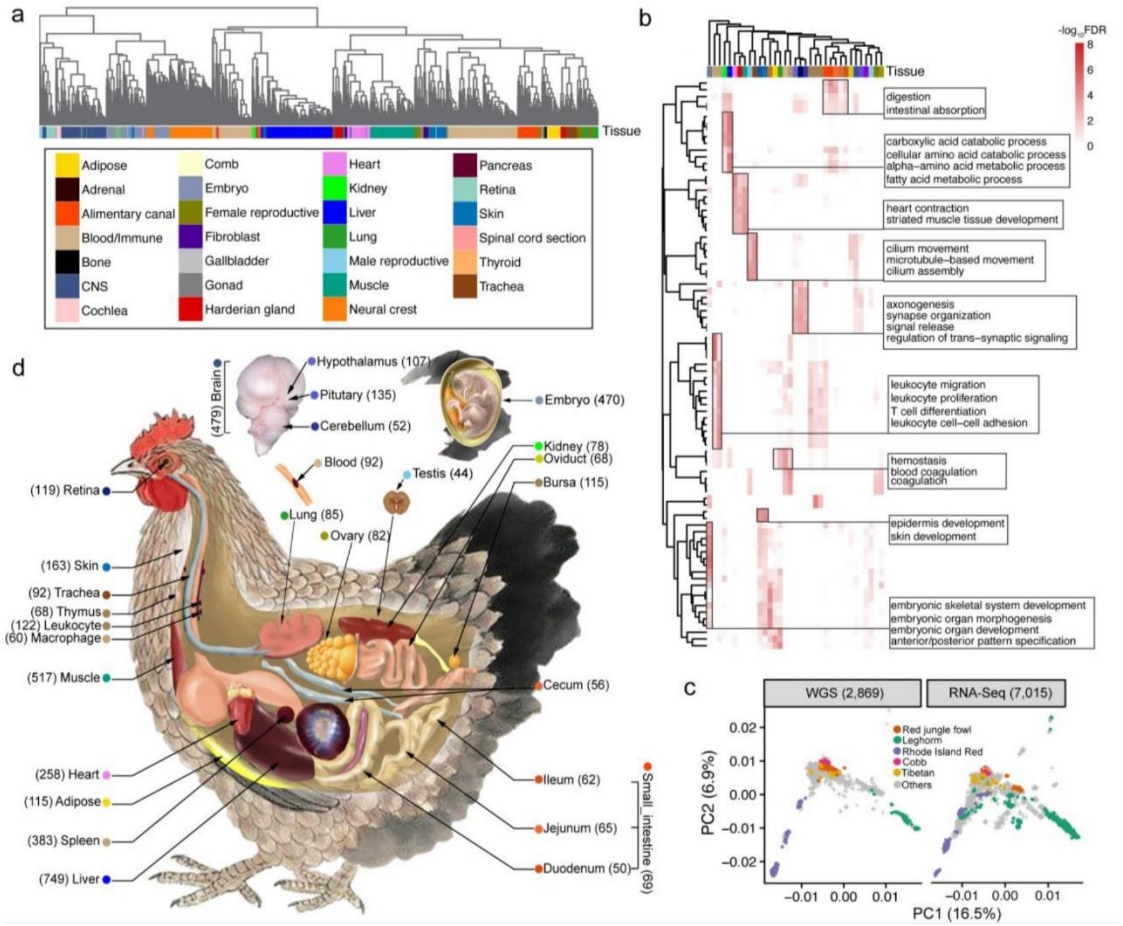


Fig. 1. Data summary in the pilot phase of ChickenGTEx. (**a**) Hierarchical clustering of 7,015 RNA-Seq samples. Distance between samples was calculated using 1-*r*, where *r* is the Pearson correlation coefficient calculated from gene expression values (quantified as Transcripts per Million, TPM) of 5,000 genes with the highest expression variance (measured by standard deviation) across samples. (**b**) Functional enrichment of tissue-specific genes based on the Gene Ontology (GO) database. The color scale from light to deep means a negative logarithm of false discovery rate (FDR) at the base of 10, obtained by the clusterProfiler 4.0 package with default settings(*103*). (**c**) Scatterplots depicting principal component analysis (PCA) of 2,869 whole-genome sequence (WGS, left) and 7,015 RNA-Seq samples (right). PCA was carried out using 1.52 million SNP genotypes shared by both WGS and RNA-Seq datasets. (**d**) Illustration of tissue types used in molecular quantitative trait loci (molQTL) mapping. Sample sizes (in the bracket) and colors of all the 28 tissues with sample sizes over 40 are depicted.

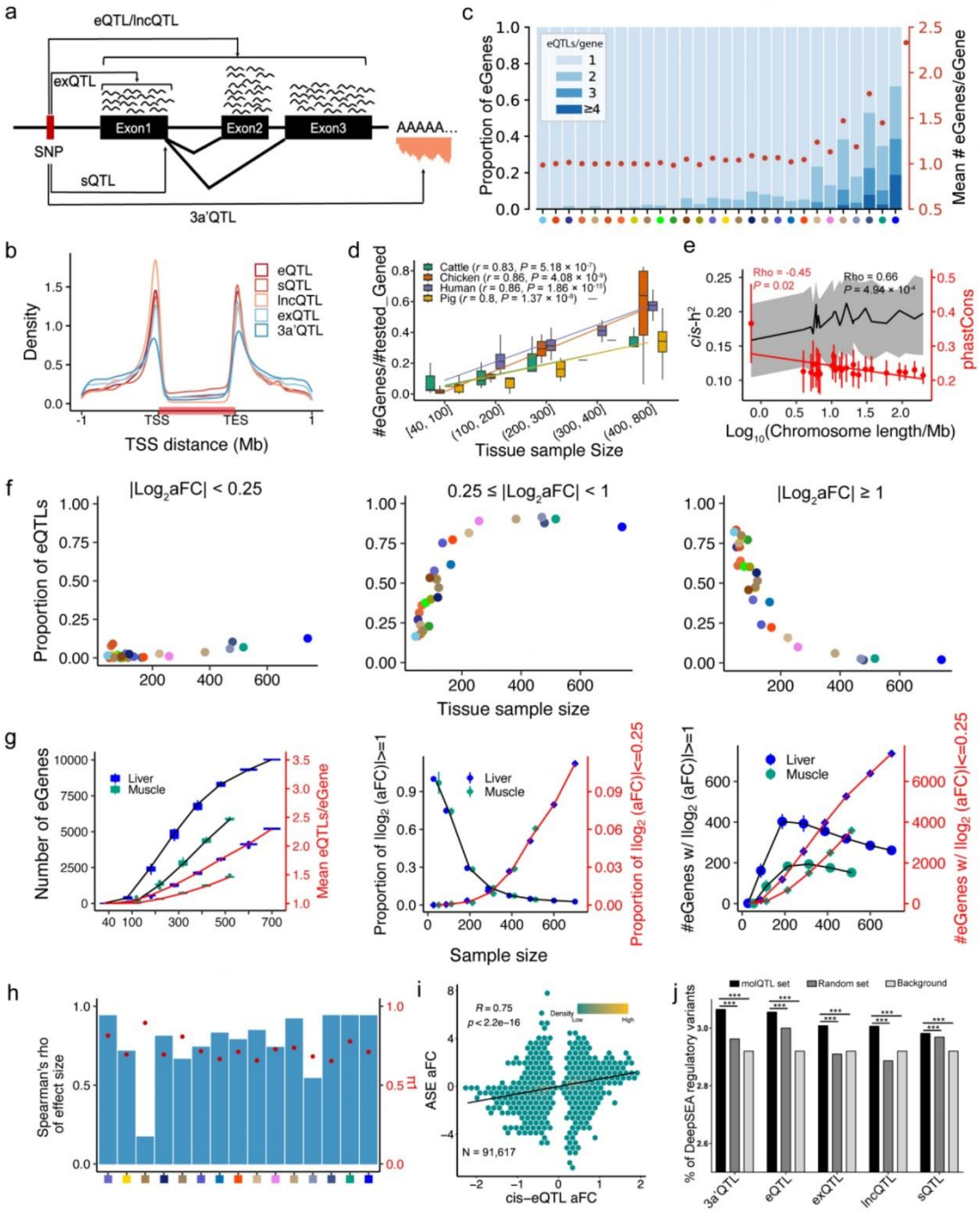
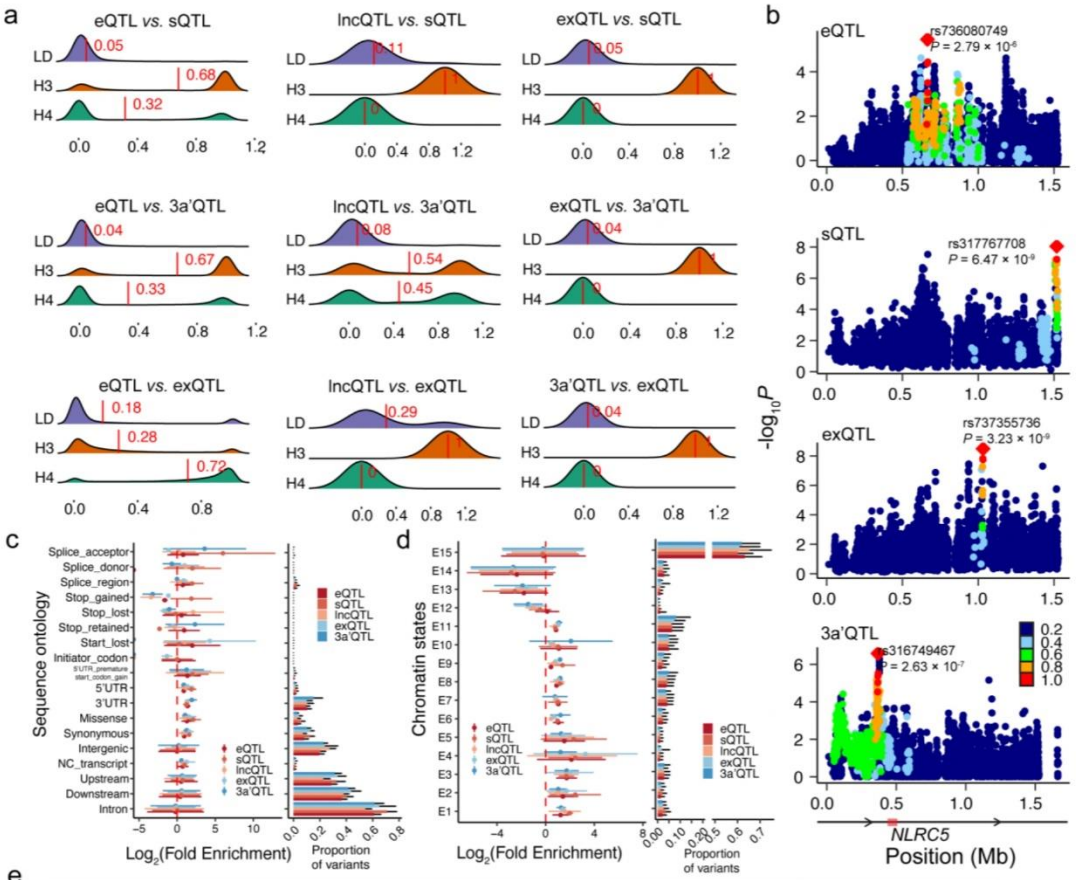


Fig. 2. Molecular QTL (molQTL) mapping in 28 chicken tissues. (**a**) Illustration of the definition of five molecular phenotypes and the respective molQTL, including protein-coding gene expression (eQTL), lncRNA expression (lncQTL), exon expression (exQTL), splicing variation (sQTL), and 3' untranslated region alternative polyadenylation (3'UTR APA, 3a'QTL). (**b**) Distribution of molQTL around gene body of eGenes, denoted by the horizontal red bar. TSS: Transcription Start Site, and TES: Transcription End Site. (**c**) Conditionally independent

eQTL across all 28 tissues. Proportion of eGenes with different numbers of independent eQTL being detected (blue stacked bars; left *y*-axis), and mean number of independent eQTL per eGene (red dots; right *y*-axis). Tissues are sorted from smallest to largest regarding sample size. Tissue color legend can be found in Fig. 1c and Table S2. (**d**) Proportion of eGenes detected as a function of tissue sample size across species, including 28, 34, 24 and 49 tissues in chickens, pigs, cattle and human, respectively. The lines are fitted with a linear model implemented in the geom_smooth function of the ggplot2 package(*104*). Correlations and *P* values were computed with the Spearman method using the cor.test function in R v3.6.5(*105*). (**e**) *cis*-h^2 (cis-heritability, left *y*-axis) and phastCons scores (right *y*-axis) of lead eQTL as a function of chromosome size ($\log_{10}$scaled). The top and bottom boundaries of the grey shade indicate the 25% and 75% of *cis*-h^2 range, respectively, and the black line is the median of *cis*-h^2 values. Red dots are average phastCons of lead eQTL, and red bars are their standard deviations. The correlations were computed with the Spearman method, and *P* values were computed *via* the asymptotic *t* approximation. (**f**) The proportion of eQTL detected (*y*-axis) with different effect sizes (from left to right panels) as a function of tissue sample size (*x*-axis). (**g**) Down-sampling analyses of eGene and eQTL. We carried out down-sampling analyses (10 replications at each sample size) in the liver and muscle, which have the largest sample size among all the 28 tissues. The left panel depicts the number of eGenes (left *y*-axis) and mean eQTL per eGene (right *y*-axis) detected at different sample size. The middle panel shows the proportion of detected eQTL of large (absolute $\log_2$aFC ≥ 1, left *y*-axis) and small effect size (absolute $\log_2$aFC ≤ 0.25, right *y*-axis). The right panel presents the number eGenes detected when the regulatory effect size of lead eQTL is large (absolute $\log_2$aFC ≥ 1, left *y*-axis) and small (absolute $\log_2$aFC ≤ 0.25, right *y*-axis). (**h**) Internal validation of eQTL. Bars in light blue indicate the Spearman correlation coefficient of eQTL effect size between validation and discovery groups (left *y*-axis), and red dots represent π_1 statistic estimating the replication rate of eQTL between groups (right *y*-axis). The samples in each of the 15 tissues with over 100 individuals are evenly and randomly divided into two groups, i.e., discovery and validation groups. The tissue color legend (*x*-axis) can be found in **Fig. 1c** and **Table S2**. (**i**) Correlation between effect size of eQTL (*x*-axis, n=91,617) and those of same loci derived from allele-specific expression (ASE, *y*-axis) analysis in liver. (**j**) The proportion of regulatory variants predicted by DeepSEA (prediction score > 0.7) based on 310 functional profiles in chickens. molQTL_set: conditionally independent molQTL across tissues; Random_set: randomly selected variants with the same MAF as molQTL; Background: all tested 1.5 million variants.



	ATAC	CTCF	DNaseSeq	H3K27ac	H3K27me3	H3K4me1	H3K4me3
3a'QTL	0.86	1.23	0.96	0.70	1.08	1.03	0.80
eQTL	0.98	0.88	0.92	1.35	1.18	0.66	1.31
exQTL	0.99	0.99	0.95	1.12	1.13	0.88	1.08
lncQTL	1.08	1.00	1.08	1.29	1.54	1.27	0.97
sQTL	1.06	0.92	0.98	1.12	1.41	0.59	1.17

Fig. 3. Colocalization and functional enrichment of molQTL. (**a**) Colocalization analyses of different types of molQTL of the same genes. The "LD" is the linkage disequilibrium (LD) of lead SNPs of two molecular phenotypes. "H3" and "H4" represent the probability of whether the association of two molecular phenotypes is due to two independent SNPs or one shared SNP, respectively. The vertical red lines indicate corresponding mean values. (**b**) Associations (i.e., -$\log_{10}$ transformed *P*) of genetic variants with four molecular phenotypes of *NLRC5*. The panels from top to bottom represent gene expression, alternative splicing, exon expression and 3'UTR APA, respectively. Color legend represents the degree of LD between the lead SNP and the others. The proportion and enrichment of five types of molQTL across sequence ontology (i.e., variant types annotated by SnpEff software(*106*)) (**c**) and 15 chromatin states (**d**). Fold enrichment is shown as mean (dot) ± standard deviation ($\log_2$ scaled, error bar) across 28 chicken tissues. The chromatin states were retrieved from Pan et al. (2023) (*30*). (**e**) The enrichment fold (odds ratio, OR) of molQTL in regulatory variants of seven epigenomic marks predicted by DeepSEA(*107*) (prediction score > 0.7). OR = (A/B)/(C/D), where C is the length of molQTL overlapped with annotated features (A), and B is the length of molQTL overlapped with the total genome length (D).

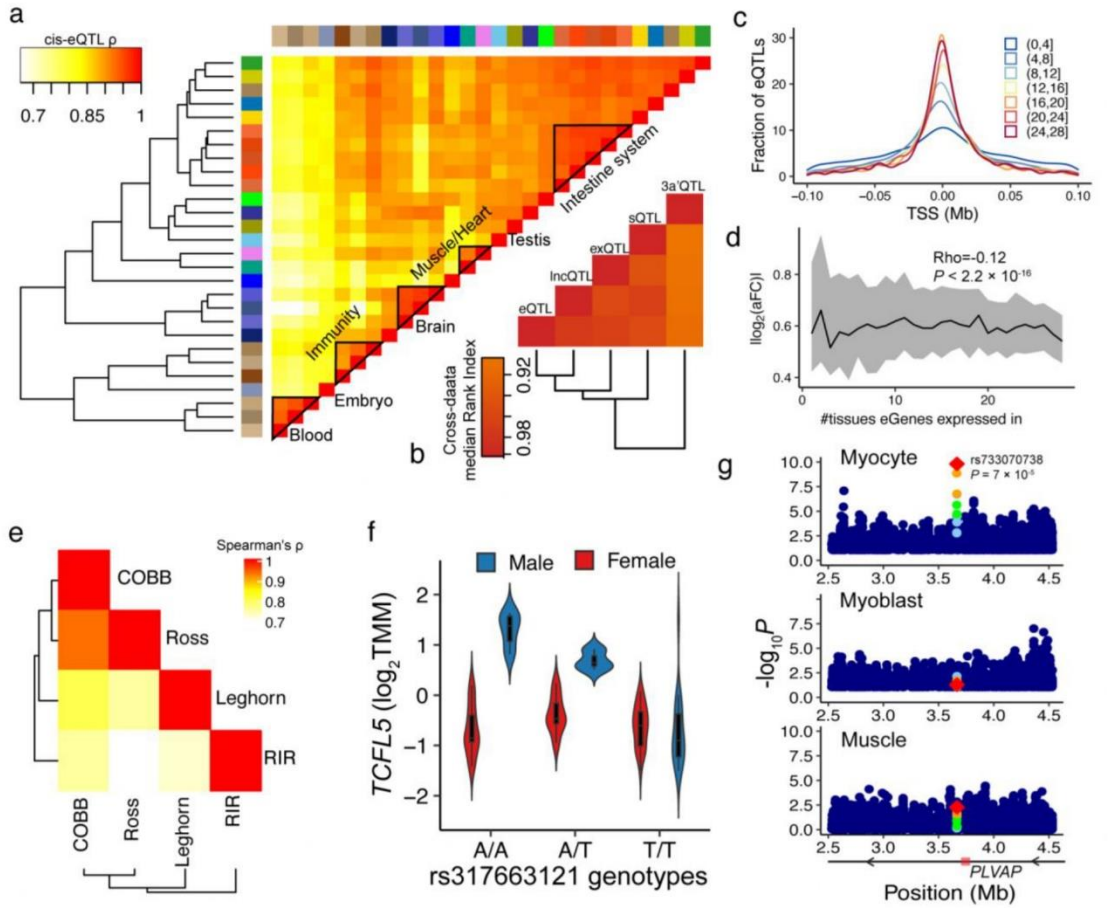


Fig. 4. Tissue-sharing and context-dependent patterns of molQTL. (**a**) The heatmap of Spearman's correlation of eQTL effect size between tissues. Tissues are clustered on the basis of dissimilarities (i.e. 1-d), where d is Euclidean distance calculated from the eQTL effect, with a complete linkage method(*108*). The color legend of tissues is the same as in **Fig. 1c** and **Table S2.** (**b**) Similarity of molQTL effect-based tissue clustering patterns. The pairwise Rand Index across five types of molQTL is used for measuring the similarity, ranging from 0 to 1, where 0 means that two tissue clustering patterns do not match at all, while 1 means that two clustering patterns match exactly. (**c**) Fraction of eQTL around transcription start site (TSS) according to number of tissues they are active in. (**d**) Absolute effect size (allelic fold change, aFC) of eQTL as a function of number of tissues where the eGene is expressed in. The black line is corresponding median estimates, and the grey shades indicate corresponding interquartile ranges. Correlation tests were carried out using *cor.test* function in R v3.6.3. (**e**) Heatmap depicting of eQTL effect sharing between breeds. This analysis was done by using Multivariate Adaptive Shrinkage (*109*), same as in panel **a**. (**f**) The expression of the *TCFL5* gene across three genotypes of *rs317663121* in males (AA, n=107; AT, n=106; TT, n=92) and female (AA, n=66; AT, n=56; TT, n=73). TMM is the Trimmed Means of M values, representing the normalized gene expression level. (**g**) The significance ($-\log_{10}P$) of interaction between *rs733070738* genotypes and myocyte enrichment (top panel) and myoblast enrichment (middle panel) on

PLVAP expression in muscle. The bottom panel is the Manhattan plot for eQTL mapping of *PLVAP* in bulk muscle samples. Dot color means linkage disequilibrium (LD) between *rs733070738* and the rest.

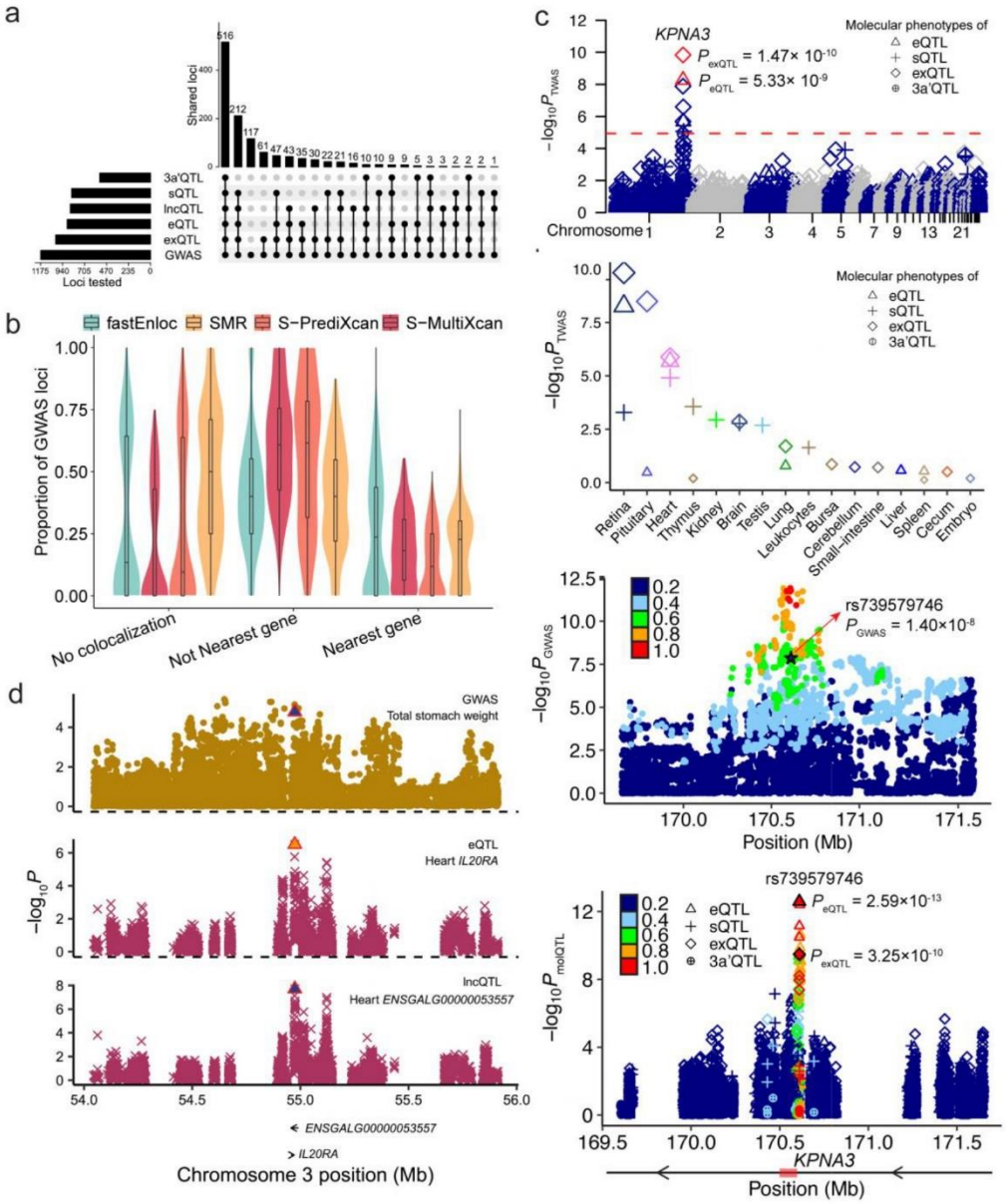


Fig. 5. Interpretation of GWAS loci with molQTL. (**a**) UpsetR plot depicting the number of GWAS loci explained by five types of molQTL, which were detected by at least one of four complementary integrative methods, including fastENLOC-based colocalization, Summary-based Mendelian randomization (SMR), single tissue-based transcriptome-wide association study (sTWAS), and multi-tissue TWAS (mTWAS). (**b**) The proportion of three types of GWAS loci (n = 1,155) colocalizing with eQTL regarding the integration results using 4 methods, including fastENLOC-based colocalization, Summary-based Mendelian randomization (SMR), single tissue-based transcriptome-wide association study (sTWAS), and multi-tissue TWAS (mTWAS). No colocalization: GWAS loci that are not interpreted by any eGenes in 28 tissues. Not nearest gene: GWAS loci are interpreted by eGenes that are not nearest genes to GWAS lead SNPs. Nearest gene: GWAS loci are interpreted by eGenes that are the nearest ones to GWAS lead SNPs. Each dot represents one of 108 complex traits. (**c**) Interpretation of GWAS loci of weight gain from week 6 to 8 (WG6.8) with molQTL. The top panel depicts associations of genes with WG6.8 *via* sTWAS in retina. The second lower panel displays associations ($-\log_{10}P$) of different molecular phenotypes (gene expression, exon expression, alternative splicing and 3'UTR APA) of *KPNA3* with WG6.8 obtained by sTWAS across tissues. The following Manhattan plot exhibits GWAS associations of SNPs with WG6.8 on chromosome 1. The color indicated linkage disequilibrium (LD) of SNPs with the lead one (*rs15497848*, $P = 1.22 \times 10^{-12}$). The colocalized SNP (*rs739579746*, $P = 1.4 \times 10^{-8}$) is denoted as a black star. The bottom plot represents molQTL mapping results of *KPNA3* in retina. The color represents LD values of the colocalized SNP (*rs739579746*, black color) with the rest. (**d**) SMR-multi results of GWAS loci of total stomach weight and eQTL and lncQTL. The top panel depicts GWAS associations of SNPs (represented by dots) with total stomach weight. The middle panel exhibits SMR associations of GWAS loci with eQTL of *IL20RA* in heart, while the bottom panel exhibits SMR associations of GWAS loci with lncQTL of *ENSGALG00000053557*. The triangle shape shows the potential causal SNP (*rs314997637*) across the three biological layers, i.e., expression of *ENSGALG00000053557,* expression of *IL20RA* and total stomach weight.

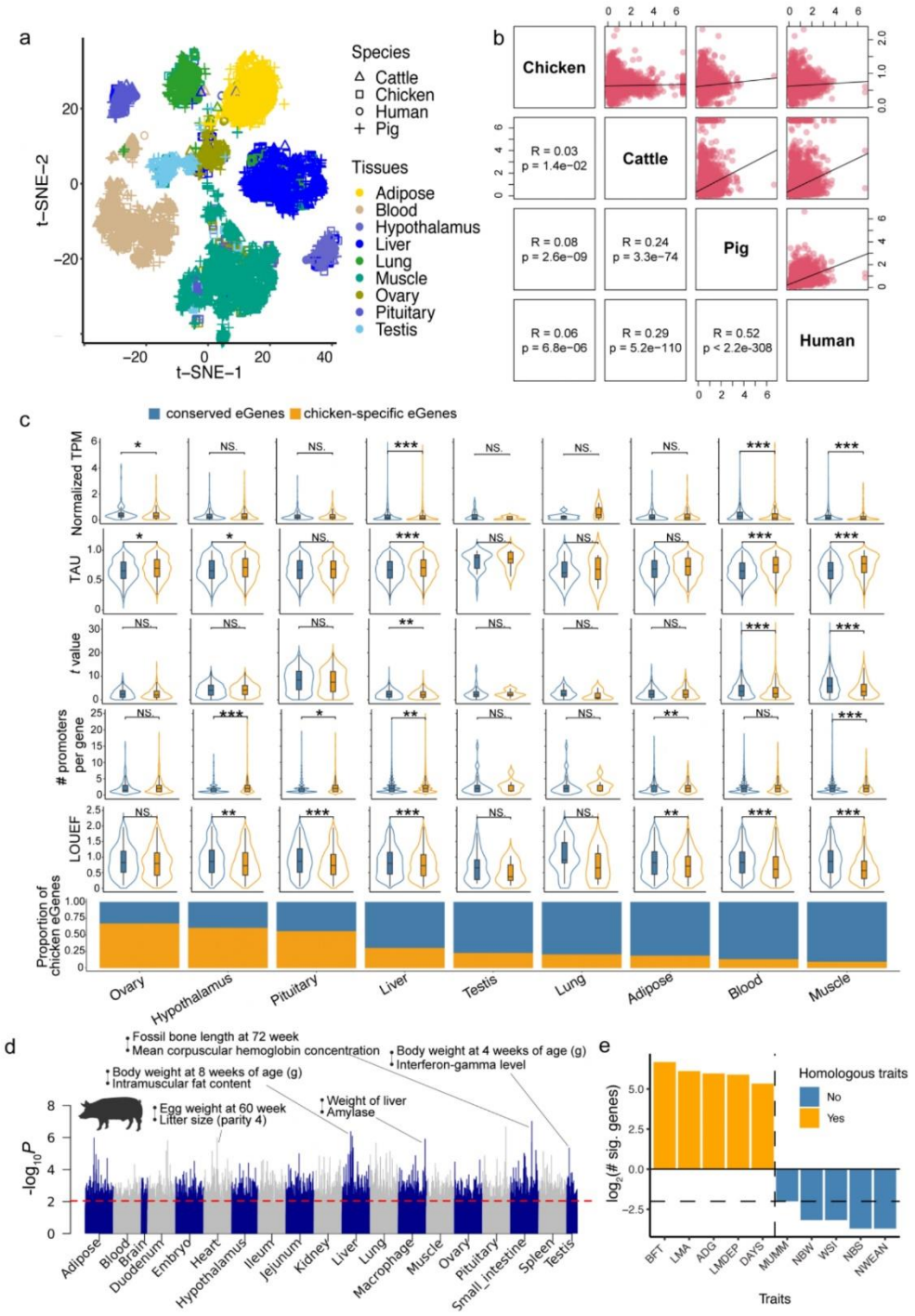


Fig. 6. Comparative analyses of gene regulation and transcriptome-wide associations (TWAS) between chickens and mammals. (**a**) Visualization of variance in gene expression of 14,278 RNA-Seq samples across four species (i.e., Chicken, pig, cattle and human) *via* a *t*-distributed stochastic neighbor embedding (*t*-SNE). Gene expression (Transcripts per Million,

TPM) of 10600 1-1-1-1 orthologous protein-coding genes are normalized between samples using Seurat software (v4.3.0) (*110*). (**b**) Pearson's correlation of averaged effect size of lead eVariants of 5,513 orthologous eGenes between species. (**c**) Comparison of chicken-specific eGenes and conserved eGenes (i.e., eGenes that are shared with at least one mammalian species) across nine tissues. The bottom barplot depicts the proportion of chicken-specific eGenes and conserved eGenes in each tissue. The violin plots from top to bottom depict expression level, TAU (tissue-specificity of expression), *t*-value (measuring the degree of gene differential expression between species), number of promoters per gene, and loss-of-function intolerance (quantified by LOEUF), respectively. Statistical tests were done by employing two-sided Wilcox-test. $^{***}P \leq 0.001$; $^{**}0.001 < P \leq 0.01$; $^{*}0.01 < P \leq 0.05$; NS: not significant ($P > 0.05$). (**d**) Significance (at $\log_{10}$ transformed, *y*-axis) for TWAS-based correlations calculated using one-to-one orthologous gene effect between chicken and pig. The red dashed line depicts the threshold of significance (Permutation-based *P* value < 0.01, corresponding to nominal *P* value < 9.11×10^{-3}). (**e**) The number of genes newly detected (FDR < 0.05) for body weight in chickens by using cross-species meta-TWAS analysis in muscle. The dashed horizontal line indicates 0 before the log transformation. Orange and blue bars represent homologous and nonhomologous traits in pigs for chicken body weight, respectively. BFT: backfat thickness, LMA: loin muscle area, ADG: average daily gain, LMDEP: loin muscle depth, DAYS: days, MUMM: number of mummified pigs, NBW: number of weak pigs, WSI: weaning to estrus interval, NBS: number of stillborn pigs, NWEAN: number of weaned piglets.

Methods and Materials

RNA-Seq data analyses and molecular phenotype definition

We downloaded 8,338 RNA-Seq data sets from the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra)and 140 public data sets from the Genome Sequence Archive (GSA, https://ngdc.cncb.ac.cn/gsa/). We also included 155 newly-generated RNA-Seq data sets. The metadata relating to all the RNA-Seq samples is summarized in **Table S1**. For quality control, we removed adaptors and trimmed low-quality reads using Trim Galore (v0.6.6, https://github.com/FelixKrueger/TrimGalore) with options of "--gzip --trim-n --length 30 --clip_R1 3 --clip_R2 3 --three_prime_clip_R1 3 --three_prime_clip_R2 3". We aligned the clean reads to the GRCg6a reference genome (Ensembl version 102) using STAR (v2.7.7a)(*111*) with parameters of "--quantMode GeneCounts --chimSegmentMin 10 --chimOutType Junctions --chimOutJunctionFormat 1 --outFilterMismatchNmax 3". For downstream analyses, only 7,015 samples with uniquely mapping rates ≥ 60 % and a number of clean reads > 500,000 after removing potentially mislabeled samples were kept. For each of these samples, we then obtained raw read counts and normalized expression (i.e., Transcripts Per Million, TPM) of 16,779 PCGs annotated in the Ensembl v102 and 22,792 lncRNA genes annotated by FR-AgENCODE (http://www.fragencode.org/)(*112*), using featureCounts (v2.0.1)(*113*) and StringTie (v2.1.5)(*114*), respectively. Using the same software(*113*), we counted the total number of reads as a function of annotated exons, which were further transformed into TPM using TBtools(*115*). We performed the tree clustering of all the RNA-Seq samples using the GGTREE package(*116*). The distance between samples was measured by 1-*r*, where *r* was Pearson's correlation coefficient based on the $\log_2$(TPM+0.25) of 5,000 genes with the highest variability. We also visualized these samples using the *t*-distributed stochastic neighbor embedding (*t*-SNE) approach implemented in the Rtsne package(*117*).

We quantified alternative splicing variation from RNA-Seq data using the LeafCutter package(*118*), which took into account spliced reads so that both novel and known alternative splicing events could be identified and quantified(*118*). Briefly, based on the STAR alignments mentioned above, we extracted junctions and defined intron clusters across samples using the script "bam2junc.sh" and leafcutter_cluster.py", respectively, as provided by the LeafCutter package(*118*). For intron clustering, we required at least 30 split reads supporting each cluster and at least 0.1% of reads supporting a junction in a cluster, as well as allowing intron length of up to 500kb. The generated matrix of per individual counts was normalized and used for clustering samples based on 1-*r*, where *r* is the Pearson's correlation coefficient between samples. To link intron clusters to genes, we mapped their coordinates to the gene model provided by the FR-AgENCODE database (*112*) using the script "map_clusters_to_genes.R" (https://github.com/broadinstitute/gtex-pipeline). Afterward, we filtered out introns if no reads were detected in >50% of samples or the number of counts was less than max(10, 0.1n) where n is the sample size. In addition, we discarded introns with low variability across samples: $\sum_i(|z_i| < 0.25) \geq$ n-3 and $\sum_i(|z_i| > 6) \leq 3$, where z_i is the z-score of the ^{i}th cluster read fraction across individuals. The filtered counts were further normalized between samples using the script "prepare_phenotype_table.py" in the LeafCutter package(*118*). The generated normalized splicing counts were stored in BED formatted file for subsequent sQTL mapping.

For the quantification of 3'UTR APA, we utilized the DaPars (v2)(*119*). We first extracted distal polyadenylation sites based on the Ensembl annotation (v102) using the script "DaPars_Extract_Anno.py". Then, we computed the genome coverage of STAR alignments mentioned above using the *genomecov* function in the BEDTools (v2.30.0)(*120*). The generated wiggle alignment files were then used for quantifying APA usage, resulting in the percentage of distal poly(A) site usage index (PDUI) value for each gene in each sample. We rescaled the PDUI values across samples to the mean of zero and variance of one in each tissue for 3a'QTL mapping.

Single-cell RNA-Seq data analyses

We retrieved single-cell RNA-Seq data from the chicken heart (n = 7) (*121*) and muscle (n = 2) (*122*) from the public database. Raw sequencing data was processed by using the "count" function after preparing the genome annotation .gtf file (Ensembl v102) with the *mkgtf* tool of the Cell Ranger pipeline(*123*). The Seurat R package (v4.0.5)(*124*) was used for subsequent cell-type identification. We first created the Seurat object based on the raw read count of each sample in a tissue using the *CreateSeuratObject* function. In this step, we filtered out cells with unique gene counts < 200 and with mitochondrial counts > 20% of the total counts. We then normalized raw counts of gene expression using the *LogNormalize* algorithm and further identified highly variable genes (HVG) using the *FindVariableFeatures* algorithm with default parameters. The HVG count matrices of all samples for a given tissue were integrated and combined to form a single *Seurat* object using the *FindIntegrationAnchors* and *IntegrateData* functions. We scaled the integrated dataset using the *ScaleData* function, which was further used to run principal components analysis (PCA) with the *RunPCA* function. The top 15 PCs, where the percentage of variance explained tended to be constant based on the elbow plot by the *JackStraw* function, were selected for running Uniform Manifold Approximation and Projection (UMAP) analysis for cell clustering using the *RunUMAP* function. The nearest neighbors between cells were constructed using the *FindNeighbors* function and cell clusters were thus determined using the *FindClusters* function at a resolution of 0.05. We manually assigned cell names based on original

894 publications (*121*, *122*) and the PanglaoDB database(*125*). Finally, cell clusters were visualized
895 using the UMAP algorithm with the *DimPlot* function. To further deconvolute bulk RNA-Seq
896 data using single-cell RNA-Seq data, we first created a signature matrix using the CIBERSORTx
897 tool (*126*) with default parameters. Using the "Impute Cell Fractions" from the same tool, we
898 imputed cell fractions with the custom mode and 1000 permutations after uploading the gene
899 expression matrix for the bulk RNA-Seq data.

900

901 **Tissue-specificity of gene expression**

902 We employed tspex(*127*), a tissue-specificity calculator, to compute 12 tissue-specificity metrics,
903 including 8 general scoring metrics (i.e. Couts, Tau, Gini coefficient, Simpson index, Shannon
904 entropy specificity, ROKU specificity, Specificity measure dispersion, and Jensen-Shannon
905 specificity dispersion) and 4 individualized scoring metrics (i.e. Tissue-specificity index, Z-
906 score, Specificity measure, and Jensen-Shannon specificity). To identify tissue-specifically
907 expressed genes in a tissue, we applied another *t*-statistic approach as described previously(*128*).
908 Briefly, for a given tissue, we carried out differential gene expression analysis between the target
909 tissue and the rest but excluded those from the same biological system using the limma
910 package(*129*). Subsequently, tissue-specific genes were identified when FDR corrected *P*-
911 value(*130*) > 0.05 and fold change > 2. Functional enrichment analysis of tissue-specific genes
912 with Biological Process (BP) terms in the Gene Ontology (GO) database was performed using
913 the clusterProfiler package(*103*).

914

915 **Sex-biased gene expression**

916 To identify genes with sex-biased expression, we employed DESeq2 software(*131*) to carry out
917 differential expression analysis between male and female samples in 18 tissues, where sample
918 size of each sex was > 10. We fitted a generalized linear model for the differential expression
919 analysis while correcting for factors, including BioProject, year when RNA-Seq data was
920 generated, age, breed, sequencing platform, library layout and selection method. After multiple
921 testing correction by the FDR approach(*130*), the set of differentially expressed genes was
922 identified when FDR corrected *P*-value < 0.01.

923

924 **Reference-guided transcript assembly**

925 Based on STAR alignment files, we assembled transcripts with the guidance of the Ensembl
926 annotation (GRCg6a v102) using the StringTie2 software tool(*114*). To increase the
927 computational efficiency, transcript assembly was run by tissue. Then, the generated assembly
928 files from all tissues were merged by using the "merge" function of the StringTie2
929 software(*114*). After quantifying the expression of assembled transcripts, we only retained
930 single-exon transcripts with TPM >1 in at least half of samples in a tissue, and multi-exon
931 transcripts with TPM >0.1 in at least half of samples in a tissue. Moreover, we compared our
932 prediction to Ensembl and NCBI annotations using GffCompare (version 0.11), and classified
933 them into 14 classes as described previously(*95*, *132*). The coding potential of predicted
934 transcripts was predicted by using CPP2 software(*133*), with lncRNA loci predicted using
935 FEElnc(*134*).

936

937 **Construction of gene co-expression networks**

938 To build gene co-expression networks in each tissue, we employed five complementary methods
939 with default parameters, including WGCNA (v1.69)(*135*), ICA (v1.0.2)(*136*), PEER (v1.3)(*137*),
940 MEGENA (v1.3.7)(*138*), and CEMiTool (v1.8.3)(*139*). The input gene expression values were
941 adjusted for hidden confounding factors by regressing out 10 PEER factors and 5 genotypic PCs
942 (see "**Preparation for molQTL mapping**" section). Functional enrichment analysis of gene co-
943 expression modules was conducted by using clusterProfiler (v4.0)(*103*), and the following
944 visualization was done using Gephi (v0.9.2)(*140*).

945

946 **SNP calling from RNA-Seq samples**

947 To call SNPs from RNA-Seq samples, we marked PCR duplicates in STAR alignment files and
948 split reads that contained Ns in their cigar string using *MarkDuplicates* and *SplitNCigarReads*
949 modules of the Genome Analysis Toolkit (GATK, v 4.1.9.0)(*45*), respectively. Using the
950 Ensembl dbSNP database (v102), we recalibrated base quality scores using GATK
951 *BaseRecalibrator* and *ApplyBQSR* modules. By following the best practice of germline variant
952 calling from RNA-Seq data, we detected small variants from the recalibrated alignments files,
953 which generated individual Genomic Variant Call Format (GVCF) files using the
954 *HaplotypeCaller* function of the GATK tool(*45*). Then, we carried out joint-calling of all GVCF
955 samples using the *GenotypeGVCFs* module from the GATK tool(*45*). For selecting high-quality
956 SNPs, we carried out a hard-filtering with criteria of "FS > 30.0 & QD < 2.0", resulting in a total
957 set of 12,191,306 SNPs.

958

959 **Construction of the multi-breed genotype imputation panel and genotype imputation**

960 We retrieved 1,693 public WGS data sets from SRA (n=1,213) and GSA (n = 480) databases
961 along with 1,176 additional newly generated WGS samples, resulting in a total set of 2,869 WGS
962 samples (**Table S6**). All raw sequence reads passed a uniform computational pipeline, including
963 adaptor removal, read alignment, and SNP calling. Briefly, we trimmed read adaptors and low-
964 quality reads using the Trimmomatic v0.39 software(*141*). The obtained clean reads were further
965 aligned against the Ensembl GRCg6a chicken reference genome (v102) using the MEM
966 algorithm of the Burrows-Wheeler Aligner (BWA, v0.7.17)(*142*). The alignment files in Binary
967 Alignment Map (BAM) format were sorted using SAMtools (v1.9)(*143*), and were further passed
968 for the removal of PCR duplicates using GATK (v4.1.9.0)(*45*). The obtained BAM files were
969 then used for variant discovery to generate individual GVCF files using the *HaplotypeCaller*
970 function of the GATK tool(*45*). The joint-calling of all 2,869 GVCF samples was further done
971 using the *GenotypeGVCFs* module from the GATK tool(*45*). For selecting high-quality SNPs,
972 we carried out a hard-filtering with criteria of "QD < 2.0, MQ < 40.0, FS > 60.0, SOR > 3.0,
973 MQRankSum < -12.5, and ReadPosRankSum < -8.0", resulting in a total set of 117,900,812
974 clean SNPs. To create the genotype imputation reference panel, we first filtered out multi-allelic
975 and sex chromosomal SNPs, as well as those with MAF < 0.01 and missing rate > 0.9 using
976 BCFtools v1.10.2(*144*), and then imputed missing genotypes using the Beagle 5.1 program(*145*).
977 This yielded the final reference panel consisting of 2,869 samples and 10,520,420 SNP
978 genotypes. To better impute SNPs called from RNA-Seq samples, we discarded SNPs called
979 from RNA-Seq samples with MAF < 0.05 using BCFtools v1.10.2(*144*) and further evaluated
980 the effect of missing rates decreasing from 0.9 gradually to 0.6 on imputation accuracy. This

evaluation revealed that the missing rate of 0.6 could reach >95% of imputation accuracy, yielding a set of 1.5 million SNPs for subsequent analysis. The genotype imputation was performed using the Beagle 5.1 program(*145*).

Preparation for molQTL mapping

Sample deduplication. After assigning RNA-Seq samples into 28 tissue types, we calculated the identity-by-state (IBS) distance between samples within each tissue based on the imputed SNP genotypes using PLINK v1.9(*146*). The formula of IBS calculation is as follows:

$$IBS = \frac{(IBS2 + 0.5 \times IBS1)}{(IBS0 + IBS1 + IBS2)}$$

where IBS0, IBS1 and IBS2 are the number of non-missing variants when IBS = 0, IBS = 1, and IBS = 2, respectively. If the IBS distance of a pair of samples is higher than 0.9, they were deemed as duplicated so that the samples with the higher sequencing depth were kept. The deduplication process was run until all IBS values per pair of samples were less than 0.9. Finally, a total of 28 tissues with sample sizes ranging from 44 (testis) to 741 (liver) were kept for subsequent molQTL mapping.

Principal component analysis. Within each of those 28 tissues, we first LD-pruned imputed genotypes with the option of "--indep-pairwise 200 100 0.1" using PLINK v1.9(*146*). Principal component analysis (PCA) of samples was then carried out, based on the LD-pruned genotypes using the smartpca tool of the EIGENSOFT v8.0.0 package(*147*). The top 5 principal components (PCs) were selected as covariates for heritability estimation and molQTL mapping.

Estimating PEER confounder factors. To correct for confounders and other unwanted technical or biological variations in RNA-Seq samples, we estimated the Probabilistic Estimation of Expression Residuals (PEER) in each of the tissues using the PEER software package(*137*). The top 10 PEER factors showing highly relative contributions (i.e., factor weight variance) to gene expression variation were selected for subsequent heritability estimation and molQTL mapping.

Phenotype preparation. For protein-coding genes, lncRNAs and exons, we filtered out features with TPM < 0.1 and raw read counts < 6 in > 20% of samples within a tissue. Raw read counts were normalized using the Trimmed Mean of M-value (TMM) algorithm of the edgeR package(*148*). The generated TMM matrix was then further normalized with an inverse normal transformation for subsequent molQTL mapping. For splicing and APA, the preparation of molecular phenotypes was described in the "**RNA-Seq data analysis and molecular phenotype definition**" section.

Estimating *cis*-heritability of gene expression

We leveraged the GCTA program v1.93.2(*149*) to estimate *cis*-heritability (*cis*-h^2) of molecular phenotypes by fitting a mixed linear model:

$$y = X\beta + g + \varepsilon$$

where y is a vector of phenotypic values (i.e., gene expressions) of all samples, β is a vector of corresponding coefficients of quantitative covariates X of all samples, which included 5 genotypic PCs and 10 PEER factors, g is a vector of the genetic values of SNPs around ±1Mb of the transcription start sites (TSS) of a gene, and ε is a vector of residuals. The genetic value g

followed a normal distribution with mean of 0 and variance of $\boldsymbol{A}\sigma_g^2$, where $\boldsymbol{A}$ was the genetic relationship matrix (GRM) between individuals(*150*). Thus, we can estimate σ_g^2, i.e., the variance explained by SNP genotypes (i.e., *cis*-h^2), using the restricted maximum likelihood (REML) approach(*150*, *151*) implemented in GCTA software(*149*). The *cis*-h^2 was finally defined when the significance level was lower than 5% based on the likelihood ratio test(*149*, *150*).

Molecular QTL mapping

In this study, we only intended to map *cis*-molQTL of each feature, i.e., SNPs distributed around 1 Mb upstream and downstream of the TSS of the gene, using tensorQTL v1.0.4(*152*). This utilized graphics processing units (GPUs) with the scalability to increase runtime and reduce the time cost. Initializing with the option of "--mode cis_nominal" of the tensorQTL v1.0.4(*152*), we calculated all nominal associations of all variant-molecular phenotype pairs. The permutation mode was further used for computing empirical *P*-values for a molecular phenotype using the option of "--mode cis" of the tensorQTL v1.0.4. After carrying out a multiple testing correction based on empirical beta-approximated *P*-values(*153*) using the false discovery rate (FDR) approach(*130*), we defined eGenes , i.e., genes that were significantly regulated by at least one variant (FDR < 0.05). For an eGene, the empirical *P*-value that was closest to an FDR of 0.05 was defined as the genome-wide empirical *P*-value threshold (pt), which was used for defining the gene-level significance threshold using qbeta(pt, beta_shape1, beta_shape2) in R (v3.6.3)(*105*), where beta_shape1 and beta_shape2 were computed by tensorQTL v1.0.4(*152*). The significant molQTL were tested SNPs whose nominal *P*-values were lower than the gene-level significance threshold.

Fine-mapping analysis of molQTL

We employed four strategies for fine-mapping independent variants underlying each molQTL. Firstly, we utilized the stepwise regression procedure for mapping conditionally independent molQTL, as used in other GTEx studies (*32*, *47*, *48*). This analysis was done by using the tensorQTL v1.0.4 with "--mode cis_independent" option(*152*). The conditionally independent molQTL mapping was based on the nominal associations mentioned above and ranked variants. Secondly, we fine-mapped putative causal variants for each molecular phenotype by using the "Sum of Single Effects" (SuSiE) model (v 1.0) (*46*). We calculated LD correlations between all tested SNPs of a molecular phenotype from the genotype reference panel and then fine-mapped variants using the SuSiE infinitesimal effect model. The posterior probability of 0.1 was used for identifying putative causal variants and credible sets.

Colocalization analysis between molecular phenotypes

To demonstrate whether two types of molecular phenotypes shared genetic regulatory mechanisms, we determined a set of paired molecular phenotypes that were transcribed from the same gene. We then ran the *coloc.abf* function in the coloc package(*154*), which is an Approximate Bayes Factor colocalization analysis for detecting significant genetic variants shared by two molecular phenotypes. The package computed posterior probabilities for: 1) no association with either molecular phenotype (H0); 2) association only with the first molecular

phenotype (H1); 3) association only with the second molecular phenotype (H2); 3) association with both molecular phenotype but two independent signals (H3); 4) association with both molecular phenotype and shared signals (H4). Moreover, we calculated the linkage disequilibrium (LD) of two lead SNPs for a pair of shared molecular phenotypes using PLINK v1.9(*146*).

Tissue- and breed-sharing of molQTL

Tissue-sharing of molQTL. To assess the cross-tissue sharing pattern of molQTL, we used Multivariate Adaptive Shrinkage in R (MashR, v0.2.57)(*109*) and METASOFT v2.0.0(*155*). For MashR, we used the z-score (slope/slope_se) of top molQTL for a gene as input. To run the *mash* model, we randomly selected 1 million molQTL-gene pairs from nominal associations being tested across all tissues by tensorQTL and obtained their z-score values. If there were missing z-score values, zero was filled and the corresponding standard error was set to $1e^6$. Local false sign rate (LFSR) was then computed by MashR and an LFSR of 0.05 was considered as the significance threshold to define whether a molQTL was active in a tissue. Pairwise Spearman's correlation of effect size of active molQTL was calculated to evaluate tissue similarity. For METASOFT, we combined all significant molQTL across tissues and computed the z-score as described above. We estimated the m-value, which represented the posterior probability indicating whether a molQTL effect exists in a tissue, using the Markov Chain Monte Carlo (MCMC) method(*156*). The m-value threshold was set as 0.7.

Breed-sharing eQTL analysis. We considered the brain (Leghorn, n = 78; Red Jungle Fowl, n = 46; Ross, n = 157), spleen (Leghorn, n = 74; Cobb, n = 43) and liver (Leghorn, n = 60; Cobb, n = 47; Ross, n = 101; Rhode Island Red, n = 78), tissues as they had more than two breeds with sample size > 40. For each breed, we ran eQTL mapping independently using tensorQTL software (v1.0.4). The eQTL sharing was assessed using METASOFT v2.0.0(*155*), and MashR (v0.2.57)(*109*), as well as $\pi 1$ statistic in the qvalue package(*51*, *157*). The METASOFT and MashR were run as described above, and the $\pi 1$ statistic (i.e. replication rate)(*157*) was used to assess if an eQTL detected in one breed can be replicated in another breed.

Detection of context-dependent QTL

Sex-biased eQTL. To identify eQTL that is significantly associated with gender, we focused on eight tissues that had at least 30 samples for each sex. In this study, we only considered conditionally independent eQTL identified above to reduce the computational burden. We fitted a linear model $y = g + s + s \times g + c + e$, where y is phenotypic values of gene expression; g is genotype (0 for homozygous ref, 1 for heterozygous, and 2 for homozygous alt); s is sex information (0 for female and 1 for male); c is quantitative covariates including 5 genotypic PCs and 10 PEER factors as we used in eQTL mapping, while e is for the residuals. The same parameters were also used for computing the null model but excluding the $s \times g$ term. We then calculated *P* values by comparing the linear interaction model to the null model using analysis of variance. The *lm()* function in R v3.6.3 (*105*) was used for model fitting.

Transcription factor (TF) interacting eQTL. To detect eQTL that may interact with the expression of transcription factors, we retrieved 956 putative transcription factors from the AnimalTFDB (v3.0). As was done for sex-biased QTL detection, we only considered conditionally independent eQTL but excluded eGenes that were TF. Likewise, we fitted the same

interaction model, but the interaction term was TF expression. A total of 15 quantitative covariates including 5 genotypic PCs and 10 PEER factors were also fitted in the model to control confounding factors. The significance threshold was set as FDR(*130*) corrected *P*-value < 0.01.

Cell-type interacting eQTL. We mapped cell-type interaction QTLs by fitting a linear model but included an interaction term implemented in the tensorQTL v1.0.4 (*152*): $y = g + i + g \times i + e$, where y is gene expression, i is the estimated abundance of cell types, and g is genetic effects estimated from SNPs within ±1Mb of the TSS of a gene, while e is for the residuals. To control confounding factors, we also included a total of 15 quantitative covariates including 5 genotypic PCs and 10 PEER factors as described above. We defined genes that had at least one significant SNP after carrying out a multiple testing correction on eigenMT-based *P*-values(*158*) using the FDR approach(*130*). We defined the threshold of significance as FDR < 0.01.

Breed interacting eQTL. To demonstrate breed effects on gene regulation, we ran breed interaction eQTL mapping using the tensorQTL (v1.0.4) tool (*152*) in brain, where the sample size of each breed was > 40, including Leghorn (n=78), Red Jungle Fowl (n = 46) and Ross (n = 157). This interaction eQTL mapping fitted the same model as described for "*Cell-type interaction eQTL"* while the interaction term was breed information. The breed origins were coded as 0 for Red Jungle Fowl and 1 for Leghorn/Ross. After a multiple testing correction using the FDR approach(*130*), gene-variant pairs with FDR < 0.01 were deemed as significant.

Estimating effect sizes of molQTL

We estimated the allelic fold change (aFC) of molQTL by employing the aFC (v0.3) Python script(*159*). The estimation was based on genotypes and molecular phenotypes (the same as molQTL mapping), as well as covariates including 5 genotypic PCs and 10 PEER factors. The 95% confidence interval of aFC was estimated by using the bootstrap method (--boot 100).

Allele specific expression (ASE)

We conducted a haplotype-based ASE analysis through the phASER (v1.1.1) software(*160*). To exclude genomic regions with high mapping error rates, we first computed the genome mappability using GenMap (v1.3.0)(*161*) with parameters: -K 75 -E 2. The generated blacklist was fitted to the phASER (v1.1.1) tool(*160*) to phase variants from the STAR alignment BAM and VCF files with options of "--paired_end 1 --mapq 255 --baseq 10". Using the script "phaser_expr_matrix.py", we measured gene-level haplotypic expression for all samples with default parameters. The generated haplotypic counts files for individual samples were further aggregated by tissue by using "phaser_expr_matrix.py". Finally, we used the "phaser_cis_var.py" script to estimate the effect size of eQTLs based on aggregated haplotypic counts. The correlation of ASE-level effect size (ASE aFC) and eQTL effect size (aFC estimated above) was computed using the Spearman's correlation approach in R v3.6.3 (*105*).

Replication of molQTL discovery

To assess the replication rate of molQTL discovery, we employed the π_1 statistic embedded in the qvalue package(*51*, *157*). Briefly, we randomly split RNA-Seq samples into two groups - QTL discovery and validation population, when the tissue sample size was greater than 100. We

ran QTL mapping in each group separately using tensorQTL (v1.0.4) (*152*) as described above. Based on replicated eQTL *P*-values, we calculated π_0 value that measured the overall proportion of true null hypotheses using the *pi0est* function within the qvalue package(*157*). The π_1 was thus obtained by 1- π_0.

DeepSEA model training and variant effect prediction

DeepSEA is a deep learning model initially trained for predicting variant effects in human(*50*), while in this study, it was retrained by utilizing 310 epigenomic peaks generated by the chicken FAANG consortium (*30*) and by Zhu et al. (*162*) (**Table S7**). According to sequencing type and histone marks, we categorized all 310 epigenomic peaks into groups, including ATAC, CTCF, DNaseSeq, H3K27ac, H3K23me3, H3K4me1 and H3K4me3, which were used as input for the model training using the Selene, a PyTorch-based package (*163*). Briefly, we grouped the genome into 200-bp bins and then labeled the bins according to input features. A genomic bin will be labeled 1 if half of the bin overlaps with an epigenomic peak, otherwise labeled as 0. The model was then trained based on a sequence region of 1,000 bp (i.e., input feature), where the 200-bp bin was placed at the center. We created validation and testing datasets by grouping chromosomes, specifically grouping chromosomes 8 and 9 to the training set and chromosomes 6 and 7 to the validation set. We computed the area under the receiver operating characteristic (AUROC) to evaluate the performance of the DeepSEA model. After that, we computed variant effect/score of two alleles for a given molQTL, i.e. 2 × 310 predicted chromatin variant scores, by inputting a 1000-bp sequence with the center being the Ref or Alt allele. The score is defined as the relative log fold changes of odds between predicted scores of the Ref and Alt. For each feature, SNPs with a score greater than 0.7 were identified as variants affecting the feature.

$$varaint\ score = \left|\log\frac{p(reference)}{1-p(reference)} - \log\frac{p(alternative)}{1-p(alternative)}\right|$$

Functional enrichment of molQTL

To understand the enrichment of molQTL in sequence ontology (i.e., SNP functional types annotated by SnpEff v5.0e) and regulatory elements (i.e., 15 chromatin states annotated in(*30*)), we employed the formula:

$$E = \frac{(C\ /\ A)}{(B\ /\ D)}$$

where A and D are the length of feature annotations and the total genome length, respectively. C is the length of molQTL overlapped with feature annotations, and B is the length of molQTL overlapped with the total genome length. To further uncover the regulatory mechanism of molQTL, we retrieved predicted pairs of regulatory elements-target genes from(*30*). We then overlapped them with our molQTL-regulated genes but at the same time required the molQTL to be located within regulatory elements. Moreover, we also performed the enrichment analysis of molQTL-regulated genes and HiC TAD with data retrieved from a previous study (*52*) using the SnakeHiC pipeline (https://github.com/FarmOmics/SnakeHiC).

Integrating molQTL with GWAS results

GWAS summary statistics. To investigate the regulatory mechanisms underpinning complex traits in pigs, we systematically integrated the identified molQTL with GWAS from 108 complex traits of economic importance, representing five trait domains (i.e., growth and development, carcass, egg production, feed efficiency and blood biochemical index). Detailed information for each GWAS is shown in **Table S17**. To perform the integrative analysis of GWAS and molQTL, we overlapped significant GWAS loci with the 1,522,091 SNPs were tested in the molQTL mapping analysis, resulting in 1,176 GWAS loci.

Enrichment of molQTL and trait-associated variants. To examine whether molQTL were significantly enriched among the significant GWAS loci, we applied QTLEnrich (v2) (*32*) to quantify the enrichment degree between significant molQTL and GWAS loci.

Transcriptome-wide association study (TWAS). We conducted single- and multi-tissue TWAS with S-PrediXcan(*164*) and S-MultiXcan(*165*) included in the MetaXcan (v0.6.11) family, respectively. Briefly, we trained the Nested Cross validated Elastic Net models with molecular phenotypes (i.e., PCG, lncRNA, splicing, exon, and 3a'Genes) and corresponding SNPs within the 1Mb *cis*-window of molecular phenotypes in all 28 tissues. The predictive models with cross-validated correlation $\rho > 0.1$ and prediction performance $P < 0.05$ were selected for subsequent analyses. Using the S-PrediXcan tool and trained models, we predicted gene-trait associations at the single-tissue level, i.e., single-tissue TWAS results. Further, using the S-MultiXcan tool, we integrated single-tissue predictions, generating the multiple-tissue TWAS results. After carrying out a multiple testing correction with the FDR approach(*130*), gene-trait associations with corrected-$P < 0.05$ were considered as significant.

Summary-based Mendelian Randomization (SMR). To explore the pleiotropic association between molecular phenotypes and a complex trait, we conducted a Mendelian Randomization analysis. This was done by using the SMR software (v1.3.1) (*166*), which can utilize summary-level data from GWAS and molQTL. To correctly fit the SMR software, the molQTL data generated by tensorQTL in this study was initially converted into BESD format with options of "--fastqtl-nominal-format --make-besd". We then ran the SMR test and carried out a multiple testing correction with the FDR approach(*130*). The gene-trait pairs with corrected *P*-value < 0.05 were selected and deemed as significant.

Colocalization analysis. To identify shared genetic variants between GWAS and molQTL, we conducted a colocalization analysis with fastENLOC (v2.0) (*167*). We first fine-mapped putative causal variants for each eGene by using a Bayesian multi-SNP genetic association analysis algorithm, deterministic approximation of posteriors (DAP, the current version is DAP-G, v1.0.0)(*168*, *169*). Leveraging the DAP-G (v1.0.0) (*168*, *169*) outcome, we generated a probabilistic annotation of molQTL using the "summarize_dap2enloc.pl" script. We then calculated approximate LD blocks using PLINK v1.9 (*146*) with options: --blocks no-pheno-req --blocks-max-kb 1000 --make-founders. The posterior inclusion probability (PIP) of GWAS loci was calculated for each LD block using TORUS (*170*) with the options: --load_zval -dump_pip. By integrating GWAS PIP values, we ran the final colocalization analysis with the fastENLOC (v2.0) tool (*171*) and obtained the regional colocalization probability (GRCP). The GRCP > 0.1 was defined as the threshold of significance.

Enrichment analysis of eQTL in selective sweeps

To determine whether domestication could be acting on regulatory variants, we retrieved selective sweeps measured by locus-specific branch length (LSBL) statistics (*14*, *65*). Briefly, we

first calculated F_{ST} for genomic windows with 20 consecutive SNPs between broilers (n=40) and Red Jungle Fowls (RJF, n=35) using VCFtools and also between layers (n=50) and Red Jungle Fowls (n=35). The LSBL values were further computed with the formula: LSBL = ($F_{ST(AB)}$ + $F_{ST(AC)}$ - $F_{ST(BC)}$) / 2. We deemed the top 0.1% of genomic windows ranked by LSBL values as significant. We examined whether eQTL were overrepresented in genomic windows under position selection, *i.e.,* whether genomic windows with at least one eQTL had higher LSBL values than the background, which has an equivalent number of windows as those of eQTL.

Comparative analysis of gene expression

To comparatively analyze gene expression across species, we collected gene expression and regulation data from multiple sources. Specifically, we obtained data for 15,044 samples from the Human GTEx web portal (v8) available at https://gtexportal.org. Additionally, we gathered gene expression data for 7,095 pig samples and 8,742 cattle samples from the FarmGTEx resource accessible at https://www.farmgtex.org/. Furthermore, as part of this study, we included gene expression and regulation data for 7,015 chicken samples. In this study, we focused on protein-coding genes based on the annotation of the Ensembl (v102), and we considered the genes with TPM > 0.1 as expressed. Specifically, we grouped chicken genes into "1-1-1-1 orthologous gene" (1-1 orthologous across species, n = 10600), "complex orthologous genes" ("1 to many", "many to 1" and "many to many", n=3644), "no homology" (without any homologous counterpart in mammals, n=2535). In total, 9 tissues in common across species (*i.e.,* adipose, blood, hypothalamus, liver, lung, muscle, ovary, pituitary, and testis) were included to conduct a comparative analysis of gene expression.

Gene expression (TPM matrix) retrieved for each was normalized using Seurat (v4.3.0) (*110*) to decrease the bias introduced by dynamic data across species. We evaluated transcriptome outcomes by counting the number of reads (reads = TPM × the length of genes (bp)) in each tissue. Samples were by performing dimensionality reduction on the normalized expression data (including 10,600 1-1-1-1 orthologous genes) with the *t*-SNE approach (*117*). Moreover, we explored the conservation of cis-heritability (h^2) and effect size (aFC) of lead eVariants across species. To do so, we selected 1-1-1-1 orthologous genes (n=5,384 for h^2) and eGenes related eVariants (n=5,513 for aFC) that are in common across species, and grouped genes into conserved eGenes (that have at least 1 homology with mammals) and chicken-specific eGenes (that didn't have homology with any mammal species).

Cross-species TWAS comparison

We performed comparative analyses of single-tissue TWAS results of 108 traits in chickens with 9,112, 1,032 and 6,480 single-tissue TWAS results in three mammalian species (*i.e.*, pigs(*47*), cattle(*48*) and humans(*32*)), representing 268, 43 and 135 complex traits, respectively. Within a shared tissue between two species, we computed Pearson's correlations between any pair of traits on the basis of z score (beta / standard error) estimated from one-to-one orthologous genes between two corresponding species. To define the threshold of significance, we carried out permutation analysis by randomly calculating Pearson's correlations between any two of all single-tissue TWAS 1000,000 times. The within-species correlations were then excluded, resulting in 609,861 Pearson's correlations and corresponding *P* values. We set the cutoff of significance as top 0.1% of permuted -$\log_{10}P$, corresponding to the *P* value of 9.11×10^{-3}. We thus conducted cross-species meta-TWAS analysis by combining TWAS results from different

1282 species based on orthologous genes. For meta-TWAS analysis, we applied a sample-size
1283 weighting (SSW) strategy (*172*) by calculating Z_{TWAS} as follows:

1284 $$Z_{TWAS} = \frac{\sum_{i=1}^{B} N_i Z_{TWASj}}{(\sum_{i=1}^{B} N_i^2)^{1/2}}$$

1285 where Z_{TWASj} is the z-score for *j*th gene in TWAS analysis, *i* is the species, *i.e.*, chicken, humans,
1286 pigs, and cattle, N_i is the number of individuals for *i*th species in TWAS, *B* is the number of
1287 species in metaTWAS. The effective sample size is $N_i = 4/(\frac{1}{N_{cases}} + \frac{1}{N_{controls}})$. To obtain the
1288 significance level, we calculated *P* values for each gene based on a Chi-squared distribution of z-
1289 scores (df=1) calculated before. After a multiple testing correction with the FDR method (*130*)
1290 by replacing original *P* value (TWAS) with *P* value (meta-TWAS) of orthologous genes, the
1291 threshold of significance was defined as FDR < 0.05.

1292

1293 **References and Notes**

1294 1. L. W. Hillier, W. Miller, Sequence and comparative analysis of the chicken genome provide
1295 unique perspectives on vertebrate evolution. Nature. 432, 695–716 (2004).
1296 2. D. W. Burt, Emergence of the chicken as a model organism: implications for agriculture and
1297 biology. Poultry Science. 86, 1460–1471 (2007).
1298 3. T. H. Beacon, J. R. Davie, The chicken model organism for epigenomic research. Genome.
1299 64, 476–489 (2021).
1300 4. P. Garcia, Y. Wang, J. Viallet, Z. Macek Jilkova, The chicken embryo model: a novel and
1301 relevant model for immune-based studies. Frontiers in Immunology. 12, 791081 (2021).
1302 5. D. Wright, C.-J. Rubin, A. Martinez Barrio, K. Schütz, S. Kerje, H. Brändström, A.
1303 Kindmark, P. Jensen, L. Andersson, The genetic architecture of domestication in the
1304 chicken: effects of pleiotropy and linkage. Molecular Ecology. 19, 5140–5156 (2010).
1305 6. J. Flores-Santin, W. W. Burggren, Beyond the chicken: alternative avian models for
1306 developmental physiological research. Front Physiol. 12, 712633 (2021).
1307 7. W. R. A. Brown, S. J. Hubbard, C. Tickle, S. A. Wilson, The chicken as a model for large-
1308 scale analysis of vertebrate gene function. Nat Rev Genet. 4, 87–98 (2003).
1309 8. Z. Wu, C. Bortoluzzi, M. F. L. Derks, L. Liu, M. Bosse, S. J. Hiemstra, M. A. M. Groenen,
1310 R. P. M. A. Crooijmans, Heterogeneity of a dwarf phenotype in Dutch traditional chicken
1311 breeds revealed by genomic analyses. Evolutionary Applications. 14, 1095–1108 (2021).
1312 9. M.-S. Wang, N. O. Otecko, S. Wang, D.-D. Wu, M.-M. Yang, Y.-L. Xu, R. W. Murphy, M.-
1313 S. Peng, Y.-P. Zhang, An evolutionary genomic perspective on the breeding of dwarf
1314 chickens. Molecular Biology and Evolution. 34, 3081–3088 (2017).
1315 10. M. Lillie, Z. Y. Sheng, C. F. Honaker, L. Andersson, P. B. Siegel, Ö. Carlborg, Genomic
1316 signatures of 60 years of bidirectional selection for 8-week body weight in chickens. Poultry
1317 Science. 97, 781–790 (2018).
1318 11. J. A. J. van der Eijk, M. B. Verwoolde, G. de Vries Reilingh, C. A. Jansen, T. B. Rodenburg,
1319 A. Lammers, Chicken lines divergently selected on feather pecking differ in immune
1320 characteristics. Physiology & Behavior. 212, 112680 (2019).
1321 12. Z. Huang, Z. Xu, H. Bai, Y. Huang, N. Kang, X. Ding, J. Liu, H. Luo, C. Yang, W. Chen, Q.
1322 Guo, L. Xue, X. Zhang, L. Xu, M. Chen, H. Fu, Y. Chen, Z. Yue, T. Fukagawa, S. Liu, G.
1323 Chang, L. Xu, Evolutionary analysis of a complete chicken genome. Proceedings of the
1324 National Academy of Sciences. 120, e2216641120 (2023).

13. R. A. Lawal, S. H. Martin, K. Vanmechelen, A. Vereijken, P. Silva, R. M. Al-Atiyat, R. S. Aljumaah, J. M. Mwacharo, D.-D. Wu, Y.-P. Zhang, P. M. Hocking, J. Smith, D. Wragg, O. Hanotte, The wild species genome ancestry of domestic chickens. BMC Biology. 18, 13 (2020).
14. M.-S. Wang, M. Thakur, M.-S. Peng, Y. Jiang, L. A. F. Frantz, M. Li, J.-J. Zhang, S. Wang, J. Peters, N. O. Otecko, C. Suwannapoom, X. Guo, Z.-Q. Zheng, A. Esmailizadeh, N. Y. Hirimuthugoda, H. Ashari, S. Suladari, M. S. A. Zein, S. Kusza, S. Sohrabi, H. Kharrati-Koopaee, Q.-K. Shen, L. Zeng, M.-M. Yang, Y.-J. Wu, X.-Y. Yang, X.-M. Lu, X.-Z. Jia, Q.-H. Nie, S. J. Lamont, E. Lasagna, S. Ceccobelli, H. G. T. N. Gunwardana, T. M. Senasige, S.-H. Feng, J.-F. Si, H. Zhang, J.-Q. Jin, M.-L. Li, Y.-H. Liu, H.-M. Chen, C. Ma, S.-S. Dai, A. K. F. H. Bhuiyan, M. S. Khan, G. L. L. P. Silva, T.-T. Le, O. A. Mwai, M. N. M. Ibrahim, M. Supple, B. Shapiro, O. Hanotte, G. Zhang, G. Larson, J.-L. Han, D.-D. Wu, Y.-P. Zhang, 863 genomes reveal the origin and domestication of chicken. Cell Res. 30, 693–701 (2020).
15. M.-S. Wang, J.-J. Zhang, X. Guo, M. Li, R. Meyer, H. Ashari, Z.-Q. Zheng, S. Wang, M.-S. Peng, Y. Jiang, M. Thakur, C. Suwannapoom, A. Esmailizadeh, N. Y. Hirimuthugoda, M. S. A. Zein, S. Kusza, H. Kharrati-Koopaee, L. Zeng, Y.-M. Wang, T.-T. Yin, M.-M. Yang, M.-L. Li, X.-M. Lu, E. Lasagna, S. Ceccobelli, H. G. T. N. Gunwardana, T. M. Senasig, S.-H. Feng, H. Zhang, A. K. F. H. Bhuiyan, M. S. Khan, G. L. L. P. Silva, L. T. Thuy, O. A. Mwai, M. N. M. Ibrahim, G. Zhang, K.-X. Qu, O. Hanotte, B. Shapiro, M. Bosse, D.-D. Wu, J.-L. Han, Y.-P. Zhang, Large-scale genomic analysis reveals the genetic cost of chicken domestication. BMC Biology. 19, 118 (2021).
16. C.-J. Rubin, M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallböök, F. Besnier, Ö. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, L. Andersson, Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 464, 587–591 (2010).
17. S. Qanbari, C.-J. Rubin, K. Maqbool, S. Weigend, A. Weigend, J. Geibel, S. Kerje, C. Wurmser, A. T. Peterson, I. L. B. Jr, R. Preisinger, R. Fries, H. Simianer, L. Andersson, Genetics of adaptation in modern chicken. PLOS Genetics. 15, e1007989 (2019).
18. C. Zhang, D. Lin, Y. Wang, D. Peng, H. Li, J. Fei, K. Chen, N. Yang, X. Hu, Y. Zhao, N. Li, Widespread introgression in Chinese indigenous chicken breeds from commercial broiler. Evolutionary Applications. 12, 610–621 (2019).
19. M. Y. Wu, G. W. Low, G. Forcina, H. van Grouw, B. P. Y.-H. Lee, R. R. Y. Oh, F. E. Rheindt, Historic and modern genomes unveil a domestic introgression gradient in a wild red junglefowl population. Evolutionary Applications. 13, 2300–2315 (2020).
20. M. Johnsson, E. Gering, P. Willis, S. Lopez, L. Van Dorp, G. Hellenthal, R. Henriksen, U. Friberg, D. Wright, Feralisation targets different genomic loci to domestication in the chicken. Nat Commun. 7, 12950 (2016).
21. E. Gering, M. Johnsson, P. Willis, T. Getty, D. Wright, Mixed ancestry and admixture in Kauai's feral chickens: invasion of domestic genes into ancient Red Junglefowl reservoirs. Molecular Ecology. 24, 2112–2124 (2015).
22. A. A. Gheyas, A. Vallejo-Trujillo, A. Kebede, M. Lozano-Jaramillo, T. Dessie, J. Smith, O. Hanotte, Integrated environmental and genomic analysis reveals the drivers of local adaptation in african indigenous chickens. Molecular Biology and Evolution. 38, 4268–4285 (2021).

23. Y. Zan, Z. Sheng, M. Lillie, L. Rönnegård, C. F. Honaker, P. B. Siegel, Ö. Carlborg, Artificial selection response due to polygenic adaptation from a multilocus, multiallelic genetic architecture. Molecular Biology and Evolution. 34, 2678–2689 (2017).
24. S. Shi, D. Shao, L. Yang, Q. Liang, W. Han, Q. Xue, L. Qu, L. Leng, Y. Li, X. Zhao, P. Dong, M. Walugembe, B. B. Kayang, A. P. Muhairwa, H. Zhou, H. Tong, Whole genome analyses reveal novel genes associated with chicken adaptation to tropical and frigid environments. Journal of Advanced Research. 47, 13-25 (2022).
25. Z.-L. Hu, C. A. Park, J. M. Reecy, Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Research. 50, D956–D961 (2022).
26. M. E. Goddard, B. J. Hayes, Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 10, 381–391 (2009).
27. L. Andersson, M. Georges, Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet. 5, 202–212 (2004).
28. J. Smith, J. M. Alfieri, N. Anthony, P. Arensburger, G. N. Athrey, J. Balacco, A. Balic, P. Bardou, P. Barela, Y. Bigot, H. Blackmon, P. M. Borodin, R. Carroll, M. C. Casono, M. Charles, H. Cheng, M. Chiodi, L. Cigan, L. M. Coghill, R. Crooijmans, N. Das, S. Davey, A. Davidian, F. Degalez, J. M. Dekkers, M. Derks, A. B. Diack, A. Djikeng, Y. Drechsler, A. Dyomin, O. Fedrigo, S. R. Fiddaman, G. Formenti, L. A. F. Frantz, J. E. Fulton, E. Gaginskaya, S. Galkina, R. A. Gallardo, J. Geibel, A. Gheyas, C. J. P. Godinez, A. Goodell, J. A. M. Graves, D. K. Griffin, B. Haase, J.-L. Han, O. Hanotte, L. J. Henderson, Z.-C. Hou, K. Howe, L. Huynh, E. Ilatsia, E. Jarvis, S. M. Johnson, J. Kaufman, T. Kelly, S. Kemp, C. Kern, J. H. Keroack, C. Klopp, S. Lagarrigue, S. J. Lamont, M. Lange, A. Lanke, D. M. Larkin, G. Larson, J. K. N. Layos, O. Lebrasseur, L. P. Malinovskaya, R. J. Martin, M. L. M. Cerezo, A. S. Mason, F. M. McCarthy, M. J. McGrew, J. Mountcastle, C. K. Muhonja, W. Muir, K. Muret, T. Murphy, I. Ng'ang'a, M. Nishibori, R. E. O'Connor, M. Ogugo, R. Okimoto, O. Ouko, H. R. Patel, F. Perini, M. I. Pigozzi, K. C. Potter, P. D. Price, C. Reimer, E. S. Rice, N. Rocos, T. F. Rogers, P. Saelao, J. Schauer, R. Schnabel, V. Schneider, H. Simianer, A. Smith, M. P. Stevens, K. Stiers, C. K. Tiambo, M. Tixier-Boichard, A. A. Torgasheva, A. Tracey, C. A. Tregaskes, L. Vervelde, Y. Wang, W. C. Warren, P. D. Waters, D. Webb, S. Weigend, A. Wolc, A. E. Wright, D. Wright, Z. Wu, M. Yamagata, C. Yang, Z.-T. Yin, M. C. Young, G. Zhang, B. Zhao, H. Zhou, Fourth Report on Chicken Genes and Chromosomes 2022. CGR (2023), doi:10.1159/000529376.
29. C. Kern, Y. Wang, X. Xu, Z. Pan, M. Halstead, G. Chanthavixay, P. Saelao, S. Waters, R. Xiang, A. Chamberlain, I. Korf, M. E. Delany, H. H. Cheng, J. F. Medrano, A. L. Van Eenennaam, C. K. Tuggle, C. Ernst, P. Flicek, G. Quon, P. Ross, H. Zhou, Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat Commun. 12, 1821 (2021).
30. Z. Pan, An atlas of regulatory elements in chicken: a resource for chicken genetics and genomics. Science Advances. 9, eade120 (2023).
31. F. W. Albert, L. Kruglyak, The role of regulatory variation in complex traits and disease. Nat Rev Genet. 16, 197–212 (2015).
32. The GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 369, 1318–1330 (2020).
33. The GTEx Consortium, Genetic effects on gene expression across human tissues. Nature. 550, 204–213 (2017).
34. The GTEx Consortium, The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 348, 648–660 (2015).

35. N. Kerimov, J. D. Hayhurst, K. Peikova, J. R. Manning, P. Walter, L. Kolberg, M. Samoviča, M. P. Sakthivel, I. Kuzmin, S. J. Trevanion, T. Burdett, S. Jupp, H. Parkinson, I. Papatheodorou, A. D. Yates, D. R. Zerbino, K. Alasoo, A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet. 53, 1–10 (2021).
36. M. Johnsson, K. B. Jonsson, L. Andersson, P. Jensen, D. Wright, Genetic regulation of bone metabolism in the chicken: similarities and differences to mammalian systems. PLOS Genetics. 11, e1005250 (2015).
37. A. Höglund, K. Strempfl, J. Fogelholm, D. Wright, R. Henriksen, The genetic regulation of size variation in the transcriptome of the cerebrum in the chicken and its role in domestication and brain size evolution. BMC Genomics. 21, 518 (2020).
38. C. Falker-Gieske, J. Bennewitz, J. Tetens, Structural variation and eQTL analysis in two experimental populations of chickens divergently selected for feather-pecking behavior. Neurogenetics. 24, 29–41 (2023).
39. A. C. Mott, A. Mott, S. Preuß, J. Bennewitz, J. Tetens, C. Falker-Gieske, eQTL analysis of laying hens divergently selected for feather pecking identifies KLF14 as a potential key regulator for this behavioral disorder. Frontiers in Genetics. 13, 969752 (2022).
40. A. Höglund, R. Henriksen, J. Fogelholm, A. M. Churcher, C. M. Guerrero-Bosagna, A. Martinez-Barrio, M. Johnsson, P. Jensen, D. Wright, The methylation landscape and its role in domestication and gene regulation in the chicken. Nat Ecol Evol. 4, 1713–1724 (2020).
41. S. Liu, L. Fang, The CattleGTEx atlas reveals regulatory mechanisms underlying complex traits. Nat Genet. 54, 1273–1274 (2022).
42. H. Ellegren, L. Hultin-Rosenberg, B. Brunström, L. Dencker, K. Kultima, B. Scholz, Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. BMC Biology. 5, 40 (2007).
43. F. W. Nicholas, Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. Nucleic Acids Research. 31, 275–277 (2003).
44. Z. Wang, L. Qu, J. Yao, X. Yang, G. Li, Y. Zhang, J. Li, X. Wang, J. Bai, G. Xu, X. Deng, N. Yang, C. Wu, An EAV-HP insertion in 5′ flanking region of SLCO1B3 causes blue eggshell in the chicken. PLOS Genetics. 9, e1003183 (2013).
45. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
46. Y. Zou, P. Carbonetto, G. Wang, M. Stephens, Fine-mapping from summary data with the "Sum of Single Effects" model. PLOS Genetics. 18, e1010299 (2022).
47. The FarmGTEx-PigGTEx Consortium, J. Teng, Y. Gao, H. Yin, Z. Bai, S. Liu, H. Zeng, L. Bai, Z. Cai, B. Zhao, X. Li, Z. Xu, Q. Lin, Z. Pan, W. Yang, X. Yu, D. Guan, Y. Hou, B. N. Keel, G. A. Rohrer, A. K. Lindholm-Perry, W. T. Oliver, M. Ballester, D. Crespo-Piazuelo, R. Quintanilla, O. Canela-Xandri, K. Rawlik, C. Xia, Y. Yao, Q. Zhao, W. Yao, L. Yang, H. Li, H. Zhang, W. Liao, T. Chen, P. Karlskov-Mortensen, M. Fredholm, M. Amills, A. Clop, E. Giuffra, J. Wu, X. Cai, S. Diao, X. Pan, C. Wei, J. Li, H. Cheng, S. Wang, G. Su, G. Sahana, M. S. Lund, J. C. M. Dekkers, L. Kramer, C. K. Tuggle, R. Corbett, M. A. M. Groenen, O. Madsen, M. Gòdia, D. Rocha, M. Charles, C. Li, H. Pausch, X. Hu, L. Frantz, Y. Luo, L. Lin, Z. Zhou, Z. Zhang, Z. Chen, L. Cui, R. Xiang, X. Shen, P. Li, R. Huang, G. Tang, M. Li, Y. Zhao, G. Yi, Z. Tang, J. Jiang, F. Zhao, X. Yuan, X. Liu, Y. Chen, X. Xu, S. Zhao, P. Zhao, C. Haley, H. Zhou, Q. Wang, Y. Pan, X. Ding, L. Ma, J. Li, P. Navarro, Q.

Zhang, B. Li, A. Tenesa, K. Li, G. E. Liu, Z. Zhang, L. Fang, A compendium of genetic regulatory effects across pig tissues (2022), doi:10.1101/2022.11.11.516073.

48. S. Liu, Y. Gao, O. Canela-Xandri, S. Wang, Y. Yu, W. Cai, B. Li, R. Xiang, A. J. Chamberlain, E. Pairo-Castineira, K. D'Mellow, K. Rawlik, C. Xia, Y. Yao, P. Navarro, D. Rocha, X. Li, Z. Yan, C. Li, B. D. Rosen, C. P. Van Tassell, P. M. Vanraden, S. Zhang, L. Ma, J. B. Cole, G. E. Liu, A. Tenesa, L. Fang, A multi-tissue atlas of regulatory variants in cattle. Nat Genet, 1–10 (2022).

49. E. Axelsson, M. T. Webster, N. G. C. Smith, D. W. Burt, H. Ellegren, Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. Genome Res. 15, 120–125 (2005).

50. J. Zhou, O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 12, 931–934 (2015).

51. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 100, 9440–9445 (2003).

52. D. Guan, Ying Wang, S. E. Aggrey, R. Okimoto, R. Hawken, H. Zhou, Profiling chromatin contacts at micro-scale in the chicken genome. International Plant and Animal Genome Conference 30. San Diego, USA (2022).

53. C. Robins, Y. Liu, W. Fan, D. M. Duong, J. Meigs, N. V. Harerimana, E. S. Gerasimov, E. B. Dammer, D. J. Cutler, T. G. Beach, E. M. Reiman, P. L. D. Jager, D. A. Bennett, J. J. Lah, A. P. Wingo, A. I. Levey, N. T. Seyfried, T. S. Wingo, Genetic control of the human brain proteome. The American Journal of Human Genetics. 108, 400–410 (2021).

54. X. Sun, Z. He, L. Guo, C. Wang, C. Lin, L. Ye, X. Wang, Y. Li, M. Yang, S. Liu, X. Hua, W. Wen, C. Lin, Z. Long, W. Zhang, H. Li, Y. Jian, Z. Zhu, X. Wu, H. Lin, ALG3 contributes to stemness and radioresistance through regulating glycosylation of TGF-β receptor II in breast cancer. Journal of Experimental & Clinical Cancer Research. 40, 149 (2021).

55. J. Chen, X. Zhao, H. Wang, Y. Chen, W. Wang, W. Zhou, X. Wang, J. Tang, Y. Zhao, X. Lu, S. Chen, L. Wang, C. Shen, S. Yang, Common variants in TGFBR2 and miR-518 genes are associated with hypertension in the Chinese population. American Journal of Hypertension. 27, 1268–1276 (2014).

56. G. Abou-Ezzi, T. Supakorndej, J. Zhang, B. Anthony, J. Krambs, H. Celik, D. Karpova, C. S. Craft, D. C. Link, TGF-β signaling plays an essential role in the lineage specification of mesenchymal stem/progenitor cells in fetal bone marrow. Stem Cell Reports. 13, 48–60 (2019).

57. M. H. Gouveia, A. R. Bentley, H. Leonard, K. A. C. Meeks, K. Ekoru, G. Chen, M. A. Nalls, E. M. Simonsick, E. Tarazona-Santos, M. F. Lima-Costa, A. Adeyemo, D. Shriner, C. N. Rotimi, Trans-ethnic meta-analysis identifies new loci associated with longitudinal blood pressure traits. Sci Rep. 11, 4075 (2021).

58. D. Noda, S. Itoh, Y. Watanabe, M. Inamitsu, S. Dennler, F. Itoh, S. Koike, D. Danielpour, P. ten Dijke, M. Kato, ELAC2, a putative prostate cancer susceptibility gene product, potentiates TGF-β/Smad-induced growth arrest of prostate cells. Oncogene. 25, 5591–5600 (2006).

59. H. Hu, Y.-R. Miao, L.-H. Jia, Q.-Y. Yu, Q. Zhang, A.-Y. Guo, AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Research. 47, D33–D38 (2019).

60. A. S. Sowa, E. Martin, I. M. Martins, J. Schmidt, R. Depping, J. J. Weber, F. Rother, E. Hartmann, M. Bader, O. Riess, H. Tricoire, T. Schmidt, Karyopherin α-3 is a key protein in

the pathogenesis of spinocerebellar ataxia type 3 controlling the nuclear localization of ataxin-3. Proceedings of the National Academy of Sciences. 115, E2624–E2633 (2018).

61. J. Pan, Y. Yang, B. Yang, Y. Yu, Artificial Polychromatic Light Affects Growth and Physiology in Chicks. PLOS ONE. 9, e113595 (2014).

62. J. Cao, Z. Wang, Y. Dong, Z. Zhang, J. Li, F. Li, Y. Chen, Effect of combinations of monochromatic lights on growth and productive performance of broilers. Poultry Science. 91, 3013–3018 (2012).

63. M. A. Estermann, A. T. Major, C. A. Smith, DMRT1 regulation of TOX3 modulates expansion of the gonadal steroidogenic cell lineage (2022), p. 2022.07.29.502037, doi:10.1101/2022.07.29.502037.

64. X. Zhang, J. Li, X. Wang, Y. Jie, C. Sun, J. Zheng, J. Li, N. Yang, S. Chen, ATAC-seq and RNA-seq analysis unravel the mechanism of sex differentiation and infertility in sex reversal chicken. Epigenetics & Chromatin. 16, 2 (2023).

65. M.-S. Wang, Y.-X. Huo, Y. Li, N. O. Otecko, L.-Y. Su, H.-B. Xu, S.-F. Wu, M.-S. Peng, H.-Q. Liu, L. Zeng, D. M. Irwin, Y.-G. Yao, D.-D. Wu, Y.-P. Zhang, Comparative population genomics reveals genetic basis underlying body size of domestic chickens. Journal of Molecular Cell Biology. 8, 542–552 (2016).

66. K.-Y. Park, H.-S. Hwang, K.-H. Cho, K. Han, G. E. Nam, Y. H. Kim, Y. Kwon, Y.-G. Park, Body weight fluctuation as a risk factor for type 2 diabetes: results from a nationwide cohort study. J Clin Med. 8, 950 (2019).

67. D. P. Berry, J. M. Lee, K. A. Macdonald, K. Stafford, L. Matthews, J. R. Roche, Associations among body condition score, body weight, somatic cell count, and clinical mastitis in seasonally calving dairy cattle. Journal of Dairy Science. 90, 637–648 (2007).

68. C. S. Fox, N. Heard-Costa, L. A. Cupples, J. Dupuis, R. S. Vasan, L. D. Atwood, Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. BMC Medical Genetics. 8, S18 (2007).

69. A. Rajendran, K. Vaidya, J. Mendoza, J. Bridwell-Rabb, S. S. Kamat, Functional annotation of ABHD14B, an orphan serine hydrolase enzyme. Biochemistry. 59, 183–196 (2020).

70. A. Rajendran, A. Soory, N. Khandelwal, G. Ratnaparkhi, S. S. Kamat, A multi-omics analysis reveals that the lysine deacetylase ABHD14B influences glucose metabolism in mammals. J Biol Chem. 298, 102128 (2022).

71. A. Ragvin, E. Moro, D. Fredman, P. Navratilova, Ø. Drivenes, P. G. Engström, M. E. Alonso, E. de la C. Mustienes, J. L. G. Skarmeta, M. J. Tavares, F. Casares, M. Manzanares, V. van Heyningen, A. Molven, P. R. Njølstad, F. Argenton, B. Lenhard, T. S. Becker, Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. Proceedings of the National Academy of Sciences. 107, 775–780 (2010).

72. S. C. Collins, H. W. Do, B. Hastoy, A. Hugill, J. Adam, M. V. Chibalina, J. Galvanovskis, M. Godazgar, S. Lee, M. Goldsworthy, A. Salehi, A. I. Tarasov, A. H. Rosengren, R. Cox, P. Rorsman, Increased expression of the diabetes gene SOX4 reduces insulin secretion by impaired fusion pore expansion. Diabetes. 65, 1952–1961 (2016).

73. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. Nat Biotechnol. 29, 24–26 (2011).

74. H. J. Taylor, Y.-H. Hung, N. Narisu, M. R. Erdos, M. Kanke, T. Yan, C. M. Grenko, A. J. Swift, L. L. Bonnycastle, P. Sethupathy, F. S. Collins, D. L. Taylor, Human pancreatic islet microRNAs implicated in diabetes and related traits by large-scale genetic analysis (2022), doi:10.1101/2022.04.21.489048.

75. T. Qi, Y. Wu, H. Fang, F. Zhang, S. Liu, J. Zeng, J. Yang, Genetic control of RNA splicing and its distinct role in complex trait variation. Nat Genet, 1–9 (2022).

76. L. Li, K.-L. Huang, Y. Gao, Y. Cui, G. Wang, N. D. Elrod, Y. Li, Y. E. Chen, P. Ji, F. Peng, W. K. Russell, E. J. Wagner, W. Li, An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. Nat Genet. 53, 994–1005 (2021).
77. D. W. Burt, Chicken genome: Current status and future opportunities. Genome Res. 15, 1692–1698 (2005).
78. M. Bergman, N. Ringertz, Gene expression pattern of chicken erythrocyte nuclei in heterokaryons. Journal of Cell Science. 97, 167–175 (1990).
79. C. Désert, E. Merlot, T. Zerjal, B. Bed'hom, S. Härtle, A. Le Cam, P.-F. Roux, E. Baeza, F. Gondret, M. J. Duclos, S. Lagarrigue, Transcriptomes of whole blood and PBMC in chickens. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics. 20, 1–9 (2016).
80. N. J. Connally, S. Nazeen, D. Lee, H. Shi, J. Stamatoyannopoulos, S. Chun, C. Cotsapas, C. A. Cassa, S. R. Sunyaev, The missing link between genetic association and regulatory function. eLife. 11, e74970 (2022).
81. Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease. Genome Biology. 18, 83 (2017).
82. U. Võsa, A. Claringbould, H.-J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Yazar, H. Brugge, R. Oelen, D. H. de Vries, M. G. P. van der Wijst, S. Kasela, N. Pervjakova, I. Alves, M.-J. Favé, M. Agbessi, M. W. Christiansen, R. Jansen, I. Seppälä, L. Tong, A. Teumer, K. Schramm, G. Hemani, J. Verlouw, H. Yaghootkar, R. Sönmez Flitman, A. Brown, V. Kukushkina, A. Kalnapenkis, S. Rüeger, E. Porcu, J. Kronberg, J. Kettunen, B. Lee, F. Zhang, T. Qi, J. A. Hernandez, W. Arindrarto, F. Beutner, J. Dmitrieva, M. Elansary, B. P. Fairfax, M. Georges, B. T. Heijmans, A. W. Hewitt, M. Kähönen, Y. Kim, J. C. Knight, P. Kovacs, K. Krohn, S. Li, M. Loeffler, U. M. Marigorta, H. Mei, Y. Momozawa, M. Müller-Nurasyid, M. Nauck, M. G. Nivard, B. W. J. H. Penninx, J. K. Pritchard, O. T. Raitakari, O. Rotzschke, E. P. Slagboom, C. D. A. Stehouwer, M. Stumvoll, P. Sullivan, P. A. C. 't Hoen, J. Thiery, A. Tönjes, J. van Dongen, M. van Iterson, J. H. Veldink, U. Völker, R. Warmerdam, C. Wijmenga, M. Swertz, A. Andiappan, G. W. Montgomery, S. Ripatti, M. Perola, Z. Kutalik, E. Dermitzakis, S. Bergmann, T. Frayling, J. van Meurs, H. Prokisch, H. Ahsan, B. L. Pierce, T. Lehtimäki, D. I. Boomsma, B. M. Psaty, S. A. Gharib, P. Awadalla, L. Milani, W. H. Ouwehand, K. Downes, O. Stegle, A. Battle, P. M. Visscher, J. Yang, M. Scholz, J. Powell, G. Gibson, T. Esko, L. Franke, Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 53, 1300–1310 (2021).
83. N. de Klein, E. A. Tsai, M. Vochteloo, D. Baird, Y. Huang, C.-Y. Chen, S. van Dam, R. Oelen, P. Deelen, O. B. Bakker, O. El Garwany, Z. Ouyang, E. E. Marshall, M. I. Zavodszky, W. van Rheenen, M. K. Bakker, J. Veldink, T. R. Gaunt, H. Runz, L. Franke, H.-J. Westra, Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. Nat Genet, 1–12 (2023).
84. B. Zeng, J. Bendl, R. Kosoy, J. F. Fullard, G. E. Hoffman, P. Roussos, Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. Nat Genet, 1–9 (2022).
85. J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K.

Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, H. F. Moore, The Genotype-Tissue Expression (GTEx) project. Nat Genet. 45, 580–585 (2013).

86. M. Georges, C. Charlier, B. Hayes, Harnessing genomic information for livestock improvement. Nat Rev Genet. 20, 135–156 (2019).

87. J. R. Xue, A. Mackay-Smith, K. Mouri, M. F. Garcia, M. X. Dong, J. F. Akers, M. Noble, X. Li, ZOONOMIA CONSORTIUM, K. Lindblad-Toh, E. K. Karlsson, J. P. Noonan, T. D. Capellini, K. J. Brennand, R. Tewhey, P. C. Sabeti, S. K. Reilly, The functional and evolutionary impacts of human-specific deletions in conserved elements. Science. 380, eabn2253 (2023).

88. I. M. Kaplow, A. J. Lawler, D. E. Schäffer, C. Srinivasan, M. E. Wirthlin, B. N. Phan, X. Zhang, K. Foley, K. Prasad, A. R. Brown, Z. Consortium, W. K. Meyer, A. R. Pfenning, Relating enhancer genetic variation across mammals to complex phenotypes using machine learning (2022), doi:10.1101/2022.08.26.505436.

89. G. Andrews, K. Fan, H. E. Pratt, N. Phalke, ZOONOMIA CONSORTIUM, E. K. Karlsson, K. Lindblad-Toh, S. Gazal, J. E. Moore, Z. Weng, Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. Science. 380, eabn7930 (2023).

90. F. W. Albert, L. Kruglyak, The role of regulatory variation in complex traits and disease. Nat Rev Genet. 16, 197–212 (2015).

91. G. Sella, N. H. Barton, Thinking about the evolution of complex traits in the era of genome-wide association studies. Annual Review of Genomics and Human Genetics. 20, 461–493 (2019).

92. M. Li, C. Sun, N. Xu, P. Bian, X. Tian, X. Wang, Y. Wang, X. Jia, R. Heller, M. Wang, F. Wang, X. Dai, R. Luo, Y. Guo, X. Wang, P. Yang, D. Hu, Z. Liu, W. Fu, S. Zhang, X. Li, C. Wen, F. Lan, A. Z. Siddiki, C. Suwannapoom, X. Zhao, Q. Nie, X. Hu, Y. Jiang, N. Yang, De novo assembly of 20 chicken genomes reveals the undetectable phenomenon for thousands of core genes on micro-chromosomes and sub-telomeric regions. Molecular Biology and Evolution, msac066 (2022).

93. K. Wang, H. Hu, Y. Tian, J. Li, A. Scheben, C. Zhang, Y. Li, J. Wu, L. Yang, X. Fan, G. Sun, D. Li, Y. Zhang, R. Han, R. Jiang, H. Huang, F. Yan, Y. Wang, Z. Li, G. Li, X. Liu, W. Li, D. Edwards, X. Kang, The chicken pan-genome reveals gene content variation and a promoter region deletion in IGF2BP1 affecting body size. Molecular Biology and Evolution (2021), doi:10.1093/molbev/msab231.

94. A. Rhie, S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim, C. Lee, B. J. Ko, M. Chaisson, G. L. Gedman, L. J. Cantin, F. Thibaud-Nissen, L. Haggerty, I. Bista, M. Smith, B. Haase, J. Mountcastle, S. Winkler, S. Paez, J. Howard, S. C. Vernes, T. M. Lama, F. Grutzner, W. C. Warren, C. N. Balakrishnan, D. Burt, J. M. George, M. T. Biegler, D. Iorns, A. Digby, D. Eason, B. Robertson, T. Edwards, M. Wilkinson, G. Turner, A. Meyer, A. F. Kautt, P. Franchini, H.

W. Detrich, H. Svardal, M. Wagner, G. J. P. Naylor, M. Pippel, M. Malinsky, M. Mooney, M. Simbirsky, B. T. Hannigan, T. Pesout, M. Houck, A. Misuraca, S. B. Kingan, R. Hall, Z. Kronenberg, I. Sović, C. Dunn, Z. Ning, A. Hastie, J. Lee, S. Selvaraj, R. E. Green, N. H. Putnam, I. Gut, J. Ghurye, E. Garrison, Y. Sims, J. Collins, S. Pelan, J. Torrance, A. Tracey, J. Wood, R. E. Dagnew, D. Guan, S. E. London, D. F. Clayton, C. V. Mello, S. R. Friedrich, P. V. Lovell, E. Osipova, F. O. Al-Ajli, S. Secomandi, H. Kim, C. Theofanopoulou, M. Hiller, Y. Zhou, R. S. Harris, K. D. Makova, P. Medvedev, J. Hoffman, P. Masterson, K. Clark, F. Martin, K. Howe, P. Flicek, B. P. Walenz, W. Kwak, H. Clawson, M. Diekhans, L. Nassar, B. Paten, R. H. S. Kraus, A. J. Crawford, M. T. P. Gilbert, G. Zhang, B. Venkatesh, R. W. Murphy, K.-P. Koepfli, B. Shapiro, W. E. Johnson, F. Di Palma, T. Marques-Bonet, E. C. Teeling, T. Warnow, J. M. Graves, O. A. Ryder, D. Haussler, S. J. O'Brien, J. Korlach, H. A. Lewin, K. Howe, E. W. Myers, R. Durbin, A. M. Phillippy, E. D. Jarvis, Towards complete and error-free genome assemblies of all vertebrate species. Nature. 592, 737–746 (2021).

95. D. Guan, M. M. Halstead, A. D. Islas-Trejo, D. E. Goszczynski, H. H. Cheng, P. J. Ross, H. Zhou, Prediction of transcript isoforms in 19 chicken tissues by Oxford Nanopore long-read sequencing. Frontiers in Genetics. 13, 997460 (2022).

96. R. I. Kuo, E. Tseng, L. Eory, I. R. Paton, A. L. Archibald, D. W. Burt, Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 18, 323 (2017).

97. S. Thomas, J. G. Underwood, E. Tseng, A. K. Holloway, on behalf of the Bench To Basinet CvDC Informatics Subcommittee, Long-read sequencing of chicken transcripts and identification of new transcript isoforms. PLOS ONE. 9, e94650 (2014).

98. J. Zhang, C. Nie, X. Li, X. Zhao, Y. Jia, J. Han, Y. Chen, L. Wang, X. Lv, W. Yang, K. Li, J. Zhang, Z. Ning, H. Bao, C. Zhao, J. Li, L. Qu, Comprehensive analysis of structural variants in chickens using PacBio sequencing. Frontiers in Genetics. 13, 971588 (2022).

99. S. Kojima, S. Koyama, M. Ka, Y. Saito, E. H. Parrish, M. Endo, S. Takata, M. Mizukoshi, K. Hikino, A. Takeda, A. F. Gelinas, S. M. Heaton, R. Koide, A. J. Kamada, M. Noguchi, M. Hamada, Y. Kamatani, Y. Murakawa, K. Ishigaki, Y. Nakamura, K. Ito, C. Terao, Y. Momozawa, N. F. Parrish, Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. Nat Genet, 1–13 (2023).

100. T. Wicker, J. S. Robertson, S. R. Schulze, F. A. Feltus, V. Magrini, J. A. Morrison, E. R. Mardis, R. K. Wilson, D. G. Peterson, A. H. Paterson, R. Ivarie, The repetitive landscape of the chicken genome. Genome Res. 15, 126–136 (2005).

101. S. Kim-Hellmuth, F. Aguet, M. Oliva, M. Muñoz-Aguirre, S. Kasela, V. Wucher, S. E. Castel, A. R. Hamel, A. Viñuela, A. L. Roberts, S. Mangul, X. Wen, G. Wang, A. N. Barbeira, D. Garrido-Martín, B. B. Nadel, Y. Zou, R. Bonazzola, J. Quan, A. Brown, A. Martinez-Perez, J. M. Soria, Gte. Consortium§, G. Getz, E. T. Dermitzakis, K. S. Small, M. Stephens, H. S. Xi, H. K. Im, R. Guigó, A. V. Segrè, B. E. Stranger, K. G. Ardlie, T. Lappalainen, Cell type–specific genetic regulation of gene expression across human tissues. Science. 369, 1332 (2020).

102. J. A. Morris, C. Caragine, Z. Daniloski, J. Domingo, T. Barry, L. Lu, K. Davis, M. Ziosi, D. A. Glinos, S. Hao, E. P. Mimitou, P. Smibert, K. Roeder, E. Katsevich, T. Lappalainen, N. E. Sanjana, Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. Science. 380, eadh7699 (2023).

103. T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, X. Fu, S. Liu, X. Bo, G. Yu, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation. 0 (2021), doi:10.1016/j.xinn.2021.100141.

104. H. Wickham, ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York, 2016).
105. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2022). Available at https://www.R-project.org/.
106. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 6, 80–92 (2012).
107. J. Zhou, O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 12, 931–934 (2015).
108. "Complete Linkage Method" in The Concise Encyclopedia of Statistics (Springer, New York, NY, 2008; https://doi.org/10.1007/978-0-387-32833-1_71), pp. 102–102.
109. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet. 51, 187–195 (2019).
110. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 33, 495–502 (2015).
111. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29, 15–21 (2013).
112. F. Jehl, K. Muret, M. Bernard, M. Boutin, L. Lagoutte, C. Désert, P. Dehais, D. Esquerré, H. Acloque, E. Giuffra, S. Djebali, S. Foissac, T. Derrien, F. Pitel, T. Zerjal, C. Klopp, S. Lagarrigue, An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. Sci Rep. 10, 20457 (2020).
113. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30, 923–930 (2014).
114. S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, M. Pertea, Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biology. 20, 278 (2019).
115. C. Chen, H. Chen, Y. Zhang, H. R. Thomas, M. H. Frank, Y. He, R. Xia, TBtools: An integrative toolkit developed for interactive analyses of big biological data. Molecular Plant. 13, 1194–1202 (2020).
116. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 8, 28–36 (2017).
117. L. van der Maaten, G. Hinton, Visualizing Data using t-SNE. Journal of Machine Learning Research. 9, 2579–2605 (2008).
118. Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, J. K. Pritchard, Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet. 50, 151–158 (2018).
119. Z. Xia, L. A. Donehower, T. A. Cooper, J. R. Neilson, D. A. Wheeler, E. J. Wagner, W. Li, Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat Commun. 5, 5274 (2014).
120. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 841–842 (2010).
121. J. Li, S. Xing, G. Zhao, M. Zheng, X. Yang, J. Sun, J. Wen, R. Liu, Identification of diverse cell populations in skeletal muscles and biomarkers for intramuscular fat of chicken by single-cell RNA sequencing. BMC Genomics. 21, 752 (2020).

122. M. Mantri, G. J. Scuderi, R. Abedini-Nassab, M. F. Z. Wang, D. McKellar, H. Shi, B. Grodner, J. T. Butcher, I. De Vlaminck, Spatiotemporal single-cell RNA sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. Nat Commun. 12, 1771 (2021).
123. G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, J. H. Bielas, Massively parallel digital transcriptional profiling of single cells. Nat Commun. 8, 14049 (2017).
124. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. Cell. 184, 3573-3587.e29 (2021).
125. O. Franzén, L.-M. Gan, J. L. M. Björkegren, PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019, baz046 (2019).
126. A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, A. A. Alizadeh, Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 37, 773–782 (2019).
127. A. P. Camargo, A. A. Vasconcelos, M. B. Fiamenghi, G. A. G. Pereira, M. F. Carazzolle, tspex: a tissue-specificity calculator for gene expression data. Research Square (2020), doi:10.21203/rs.3.rs-51998/v1.
128. L. Fang, W. Cai, S. Liu, O. Canela-Xandri, Y. Gao, J. Jiang, K. Rawlik, B. Li, S. G. Schroeder, B. D. Rosen, C. Li, T. S. Sonstegard, L. J. Alexander, C. P. V. Tassell, P. M. VanRaden, J. B. Cole, Y. Yu, S. Zhang, A. Tenesa, L. Ma, G. E. Liu, Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. Genome Res. 30, 790-801 (2020).
129. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 43, e47 (2015).
130. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). 57, 289–300 (1995).
131. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 15, 550 (2014).
132. G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare. F1000Research (2020), , doi:10.12688/f1000research.23297.2.
133. X. Tong, S. Liu, CPPred: coding potential prediction based on the global description of RNA sequence. Nucleic Acids Research. 47, e43 (2019).
134. V. Wucher, F. Legeai, B. Hédan, G. Rizk, L. Lagoutte, T. Leeb, V. Jagannathan, E. Cadieu, A. David, H. Lohi, S. Cirera, M. Fredholm, N. Botherel, P. A. J. Leegwater, C. Le Béguec, H. Fieten, J. Johnson, J. Alföldi, C. André, K. Lindblad-Toh, C. Hitte, T. Derrien, FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Research. 45, e57 (2017).

135. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9, 559 (2008).
136. A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications. Neural Networks. 13, 411–430 (2000).
137. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 7, 500–507 (2012).
138. W.-M. Song, B. Zhang, Multiscale embedded gene co-expression network analysis. PLOS Computational Biology. 11, e1004574 (2015).
139. P. S. T. Russo, G. R. Ferreira, L. E. Cardozo, M. C. Bürger, R. Arias-Carrasco, S. R. Maruyama, T. D. C. Hirata, D. S. Lima, F. M. Passos, K. F. Fukutani, M. Lever, J. S. Silva, V. Maracaja-Coutinho, H. I. Nakaya, CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics. 19, 56 (2018).
140. M. Leonard, S. Graham, D. Bonacum, The human factor: the critical importance of effective teamwork and communication in providing safe care. BMJ Quality & Safety. 13, i85–i90 (2004).
141. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30, 2114–2120 (2014).
142. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013), doi:10.48550/arXiv.1303.3997.
143. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25, 2078–2079 (2009).
144. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. GigaScience. 10, giab008 (2021).
145. B. L. Browning, Y. Zhou, S. R. Browning, A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics. 103, 338–348 (2018).
146. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 81, 559–575 (2007).
147. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38, 904–909 (2006).
148. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26, 139–140 (2010).
149. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics. 88, 76–82 (2011).
150. J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, P. M. Visscher, Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 42, 565–569 (2010).
151. H. D. PATTERSON, R. THOMPSON, Recovery of inter-block information when block sizes are unequal. Biometrika. 58, 545–554 (1971).

152. A. Taylor-Weiner, F. Aguet, N. J. Haradhvala, S. Gosai, S. Anand, J. Kim, K. Ardlie, E. M. Van Allen, G. Getz, Scaling computational genomics to millions of individuals with GPUs. Genome Biology. 20, 228 (2019).
153. H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, O. Delaneau, Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 32, 1479–1485 (2016).
154. C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, V. Plagnol, Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLOS Genetics. 10, e1004383 (2014).
155. D. Duong, L. Gai, S. Snir, E. Y. Kang, B. Han, J. H. Sul, E. Eskin, Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. Bioinformatics. 33, i67–i74 (2017).
156. J. S. Speagle, A conceptual introduction to Markov Chain Monte Carlo methods (2020), doi:10.48550/arXiv.1909.12313.
157. Storey JD, Bass AJ, Dabney A, Robinson D, qvalue: Q-value estimation for false discovery rate control (2022), (available at https://github.com/StoreyLab/qvalue).
158. J. R. Davis, L. Fresard, D. A. Knowles, M. Pala, C. D. Bustamante, A. Battle, S. B. Montgomery, An efficient multiple-testing adjustment for eqtl studies that accounts for linkage disequilibrium between variants. The American Journal of Human Genetics. 98, 216–224 (2016).
159. P. Mohammadi, S. E. Castel, A. A. Brown, T. Lappalainen, Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. Genome Res. (2017), doi:10.1101/gr.216747.116.
160. S. E. Castel, F. Aguet, P. Mohammadi, F. Aguet, S. Anand, K. G. Ardlie, S. Gabriel, G. A. Getz, A. Graubert, K. Hadley, R. E. Handsaker, K. H. Huang, S. Kashin, X. Li, D. G. MacArthur, S. R. Meier, J. L. Nedzel, D. T. Nguyen, A. V. Segrè, E. Todres, F. Aguet, S. Anand, K. G. Ardlie, B. Balliu, A. N. Barbeira, A. Battle, R. Bonazzola, A. Brown, C. D. Brown, S. E. Castel, D. F. Conrad, D. J. Cotter, N. Cox, S. Das, O. M. de Goede, E. T. Dermitzakis, J. Einson, B. E. Engelhardt, E. Eskin, T. Y. Eulalio, N. M. Ferraro, E. D. Flynn, L. Fresard, E. R. Gamazon, D. Garrido-Martín, N. R. Gay, G. A. Getz, M. J. Gloudemans, A. Graubert, R. Guigó, K. Hadley, A. R. Hame, R. E. Handsaker, Y. He, P. J. Hoffman, F. Hormozdiari, L. Hou, K. H. Huang, H. K. Im, B. Jo, S. Kasela, S. Kashin, M. Kellis, S. Kim-Hellmuth, A. Kwong, T. Lappalainen, X. Li, X. Li, Y. Liang, D. G. MacArthur, S. Mangul, S. R. Meier, P. Mohammadi, S. B. Montgomery, M. Muñoz-Aguirre, D. C. Nachun, J. L. Nedzel, D. T. Nguyen, A. B. Nobel, M. Oliva, Y. S. Park, Y. Park, P. Parsana, A. S. Rao, F. Reverter, J. M. Rouhana, C. Sabatti, A. Saha, A. V. Segrè, A. D. Skol, M. Stephens, B. E. Stranger, B. J. Strober, N. A. Teran, E. Todres, A. Viñuela, G. Wang, X. Wen, F. Wright, V. Wucher, Y. Zou, P. G. Ferreira, G. Li, M. Melé, E. Yeger-Lotem, M. E. Barcus, D. Bradbury, T. Krubit, J. A. McLean, L. Qi, K. Robinson, N. V. Roche, A. M. Smith, L. Sobin, D. E. Tabor, A. Undale, J. Bridge, L. E. Brigham, B. A. Foster, B. M. Gillard, R. Hasz, M. Hunter, C. Johns, M. Johnson, E. Karasik, G. Kopen, W. F. Leinweber, A. McDonald, M. T. Moser, K. Myer, K. D. Ramsey, B. Roe, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, S. D. Jewell, D. C. Rohrer, D. R. Valley, D. A. Davis, D. C. Mash, M. E. Barcus, P. A. Branton, L. Sobin, L. K. Barker, H. M. Gardiner, M. Mosavel, L. A. Siminoff, P. Flicek, M. Haeussler, T. Juettemann, W. J. Kent, C. M. Lee, C. C. Powell, K. R. Rosenbloom, M. Ruffier, D. Sheppard, K. Taylor, S. J. Trevanion, D. R. Zerbino, N. S. Abell, J. Akey, L. Chen, K. Demanelis, J. A. Doherty, A. P. Feinberg, K. D. Hansen, P. F. Hickey, L. Hou, F. Jasmine, L. Jiang, R. Kaul, M. Kellis, M. G. Kibriya, J. B. Li, Q. Li, S. Lin, S. E. Linder, S. B. Montgomery, M. Oliva, Y. Park, B. L.

Pierce, L. F. Rizzardi, A. D. Skol, K. S. Smith, M. Snyder, J. Stamatoyannopoulos, B. E. Stranger, H. Tang, M. Wang, P. A. Branton, L. J. Carithers, P. Guan, S. E. Koester, A. R. Little, H. M. Moore, C. R. Nierras, A. K. Rao, J. B. Vaught, S. Volpi, K. G. Ardlie, T. Lappalainen, GTEx Consortium, A vast resource of allelic expression data spanning human tissues. Genome Biology. 21, 234 (2020).

161. C. Pockrandt, M. Alzamel, C. S. Iliopoulos, K. Reinert, GenMap: ultra-fast computation of genome mappability. Bioinformatics. 36, 3687–3692 (2020).
162. X.-N. Zhu, Y.-Z. Wang, C. Li, H.-Y. Wu, R. Zhang, X.-X. Hu, Chicken chromatin accessibility atlas accelerates epigenetic annotation of birds and gene fine-mapping associated with growth traits. Zool Res. 44, 53–62 (2023).
163. K. M. Chen, E. M. Cofer, J. Zhou, O. G. Troyanskaya, Selene: a PyTorch-based deep learning library for sequence data. Nat Methods. 16, 315–318 (2019).
164. A. N. Barbeira, S. P. Dickinson, R. Bonazzola, J. Zheng, H. E. Wheeler, J. M. Torres, E. S. Torstenson, K. P. Shah, T. Garcia, T. L. Edwards, E. A. Stahl, L. M. Huckins, D. L. Nicolae, N. J. Cox, H. K. Im, Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 9, 1825 (2018).
165. A. N. Barbeira, M. Pividori, J. Zheng, H. E. Wheeler, D. L. Nicolae, H. K. Im, Integrating predicted transcriptome from multiple tissues improves association detection. PLOS Genetics. 15, e1007889 (2019).
166. Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, J. Yang, Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 48, 481–487 (2016).
167. X. Wen, R. Pique-Regi, F. Luca, Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLOS Genetics. 13, e1006646 (2017).
168. Y. Lee, F. Luca, R. Pique-Regi, X. Wen, Bayesian Multi-SNP genetic association analysis: Control of FDR and use of summary statistics. BioRxiv. 316471 (2018).
169. X. Wen, Y. Lee, F. Luca, R. Pique-Regi, Efficient integrative Multi-SNP association analysis using deterministic approximation of posteriors. Am J Hum Genet. 98, 1114-1129 (2016).
170. X. Wen, Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. The Annals of Applied Statistics. 10, 1619–1638 (2016).
171. A. Hukku, M. G. Sampson, F. Luca, R. Pique-Regi, X. Wen, Analyzing and reconciling colocalization and transcriptome-wide association studies from the perspective of inferential reproducibility. The American Journal of Human Genetics. 109, 825–837 (2022).
172. A. Bhattacharya, J. B. Hirbo, D. Zhou, W. Zhou, J. Zheng, M. Kanai, B. Pasaniuc, E. R. Gamazon, N. J. Cox, Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: Lessons from the Global Biobank Meta-analysis Initiative. Cell Genomics. 2 (2022), doi:10.1016/j.xgen.2022.100180.

Supplementary Materials

Materials and Methods

Figs. S1 to S32

Tables S1 to S22

3.2. Détection de régions eQTL dans le foie d'une population commerciale de poules pondeuses et lien avec des QTL liés au métabolisme des lipides *(Résumé de travaux – co-encadrement d'un stage de M2)*

3.2.1. Contexte et objectifs

Ces travaux ont été produits lors du stage de Master 2 d'Alexandre Hubert que j'ai eu l'opportunité de co-encadrer et auquel j'ai vivement participé. L'exposé des méthodes et résultats est donc en grande partie issu du rapport de stage associé.

Ce stage est inscrit dans le cadre du projet EFFICACE [537], financé par l'ANR, et du projet européen GEroNIMO [538] (voir Résultats §2.3). L'un des objectifs communs à ces deux projets est l'étude génétique de la poule pondeuse à âge avancé de production, soit 90 semaines. En raison du peu d'études scientifiques conduites à ces âges, les deux projets visent à mieux comprendre les bases génétiques responsables de la variation inter-individuelle de différents caractères d'intérêt, dont le taux de ponte à 90 semaines.

D'un point de vue biologique, la production d'œufs chez la poule nécessite une quantité importante d'énergie et en particulier de lipides qui sont un composant majeur du jaune d'œuf. Dans le cadre du stage, le foie a été choisi comme tissu d'intérêt, car il est l'organe central du métabolisme énergétique et notamment de la synthèse des lipides (acides gras) de l'organisme, les lipides hépatiques étant un constituant majeur du jaune d'œuf. Les caractères étudiés au cours du stage sont en lien avec les lipides du foie et ont la particularité d'être imbriqués, débutant par *i)* des caractères macroscopiques comme l'adiposité corporelle et le poids de foie, puis, *ii)* des caractères moléculaires comme la quantité d'acides gras dans le foie qui, si elle est trop importante, peut conduire à une stéatose [592], voire une pathologie de type fibrose et pour finir *iii)* la composition de ces acides gras en acides gras saturés, mono saturés et polyinsaturés. Pour finir, l'expression des gènes du foie, s'inscrivant à l'interface entre génotype et caractères d'intérêts, a été étudiée.

Le stage avait plusieurs objectifs, le premier consistait en la mise au point d'un *pipeline* permettant de réaliser à la fois les GWAS sur des phénotypes macroscopiques et expressionnels. Deux outils d'analyse GWAS, GEMMA et GCTA ont été comparés sur trois phénotypes macroscopiques plus ou moins directement associés au foie : le gras corporel, le

poids du foie et les quantités d'acides gras dans le foie. Le second objectif a consisté à mettre en place l'approche GWAS sur des données d'expression de gènes dans le foie par l'utilisation de l'outil TensorQTL. Pour finir, une détection des QTL sur les phénotypes macroscopiques ainsi qu'une détection des eQTL pour les gènes exprimés dans le foie a été effectué.

À la différence, des travaux présentés dans le cadre du ChickenGTEX (voir Résultats §3.1), la population est ici homogène et les covariables pouvant impacter les analyses sont en partie identifiées. De plus, les caractères mesurés, qu'ils soient macroscopiques ou expressionnels, le sont pour les mêmes individus et ont été choisies pour leurs liens forts supposés (étude de caractères lipidiques et expression du foie). L'objectif est donc d'observer l'impact d'un tel dispositif sur les analyses conjointes eQTL/QTL.

3.2.2. Matériels et démarche

Données disponibles

- *Phénotypes macroscopique lipidiques mesurés à l'échelle de l'animal*

L'étude s'intéresse à sur une cohorte de 940 poules pondeuses de 90 semaines, de race *Rhode-island Red* et provenant d'une lignée commerciale *Novo* du sélectionneur privé NOVOGEN. Notons que ces individus appartiennent à trois lots d'élevage notés 2017, 2019 et 2020 et correspondant à leurs années de naissance. Pour les 940 individus, le poids de l'animal, le poids du gras corporel et le poids de foie ont été entre autres mesurés à 90 semaines. De plus, pour 247 individus du lot 2020, les quantités d'acides gras totaux (AGT), saturées (AGS), mono-insaturés (AGMI) et poly-insaturés (AGPI) dans le foie ont été quantifiées (voir Table 1).

- *Transcriptomes hépatiques et critères d'expression*

Les transcriptomes de 363 individus répartis dans les trois lots ont été séquencés (voir Table 1) par séquençage Illumina RNAseq *paired-end*. Les données de séquençage ont été alignées sur le génome de référence GRCg7b (GCF_016699485.2) avec l'annotation (*.gtf*) enrichie en modèle génique (voir Résultats §1.1) et contenant 24 102 PCG et 44 428 lncRNA [125]. La quantification de l'expression des gènes a été traitée avec le *pipeline* « rnaseq » v3.8.1 (aligner : *star rsem*) de « nf-core » [451, 451]. L'expression normalisée TMM a quant à elle été obtenue à partir des comptages bruts en utilisant le package R edgeR. Un gène a été considéré

comme exprimé si le nombre de *reads* était ≥ 6 et si son expression était ≥ 0,1 TPM & 0,1 TMM pour au moins 75% des échantillons.

- *Génotypes des animaux*

Les 940 individus ont été génotypés par une puce à SNP 60k (voir Table 1) et les contrôles qualités ont été appliqués tels que présentés dans les Résultats §2.3. Ainsi, 42 418 des 60k SNP de départ présentent un polymorphisme dans cette population.

Table 1 – Effectifs des différents phénotypes étudiés

	Lot de 2020 (349)	**Lots de 2019** (344)	**Lot de 2017 (247)**
Age	90 semaines	90 semaines	
Phénotypes			
Gras corporel	349	344	247
Poids de foie	349	344	247
Acides gras totaux	247	.	.
Phénotypes d'expression			
RNAseq	176	86	101
Génotypes			
Puce 60k	349	344	247

- *Covariables identifiées dans le plan expérimental*

Les covariables identifiées et testées dans les différents modèles (Table 2) sont :

- le poids de l'animal ;
- le poids du foie de l'animal ;
- le lot de naissance (2017, 2019, 2020) ;
- l'emplacement de la cage dans le bâtiment d'élevage, ce paramètre a été subdivisé en 32 modalités quadrillant le bâtiment ;
- son système d'élevage avant 55 semaines. En effet, 117 individus du lot 2020 ont été élevés au sol avant 55 semaines puis mis en cages individuelles par la suite, tandis que les 823 autres individus ont été élevés en cages collectives avant 55 semaines puis placés en cages individuelles ;
- l'âge exacte en jours lors de l'abattage ;
- l'effet père.

Table 2 – Covariables testées pour les différents phénotypes analysés par GWAS

Facteurs testés	Gras corporel	Poids de foie	Acides gras	Expressions
Poids de l'animal	X	X		
Poids du foie			X	X
Lot de l'animal	X	X		X
Emplacement de la cage	X	X	X	X
Sys. avant 55 sem.	X	X	X	X
Age en jours	X	X	X	X
Effet père	X	X	X	X

Applications de la méthode QTL

- *Significativité des covariables*

La significativité des covariables suspectées a été testée via l'utilisation d'un modèle linéaire mixte. Ces covariables ont été placées comme effets fixes dans le modèle. Afin de considérer la hiérarchie génétique des données, l'identifiant du père a été testé comme effet aléatoire du modèle. Une covariable est alors considérée comme significative, si la p-value qui lui est associée est < 0,2. De façon itérative, un nouveau modèle est ainsi testé, ne conservant que les covariables considérées comme significatives. Une fois le modèle établi, il a ensuite été validé en vérifiant la normalité des résidus ainsi que leur homogénéité.

- *Les modèles GWAS de GEMMA et GCTA*

Les covariables significatives ont été intégrés dans les modèles, cependant, alors que GCTA permet de traiter à la fois les covariables quantitatives et qualitatives, GEMMA ne traite que des covariables quantitatives. Ainsi, pour GEMMA, les covariables qualitatives ont été converties en covariables quantitatives en transformant les N modalités de la covariable qualitative en N covariables quantitatives contenant des 0 ou des 1.

Lors de l'utilisation de GEMMA [593], le modèle ULMM (*Univariate Linear Mixed Model*) a été utilisé. Ce modèle prend en compte l'apparentement des individus via une matrice génomique de parenté fournie en entrée. Pour GCTA [594], c'est le modèle MLM LOCO (*Mixed Linear Model – Leaving One Chromosome Out*) qui a été utilisé. Ce modèle est similaire à celui de GEMMA à la seule différence que la matrice génomique de parenté est calculée automatiquement et que durant son élaboration, les SNP présents sur le même chromosome que le SNP testé sont retirés. Dans les deux cas, le QTL est ensuite défini comme la région de plus ou moins 500 kb entourant le variant *leader*.

Applications de la méthode eQTL

- *Outil et intégration des covariables*

Les eQTL ont été détectés via l'utilisation de TensorQTL [415]. L'apparentement des individus a été pris en compte en incluant les coordonnées des cinq premiers axes d'ACP sur les génotypes des individus. Les autres covariables ont été pris en compte via les coordonnées des axes d'ACP réalisée sur les expressions des gènes, normalisées en TPM. Les corrélations entre les différents effets et les 15 premiers axes d'ACP sur les expressions ont été calculées dans le but de déterminer le nombre d'axes à considérer pour avoir une bonne représentation des covariables, ces dernières ne pouvant être incluses directement dans l'outil.

- *Critères d'identification des eGene*

Les eQTL locaux et distants correspondent à des régions d'1Mb soit 500kb de part et d'autre du *eVariant leader*. Les 15 135 gènes exprimés ont été corrigés pour les tests multiples par Bonferroni.

- *Intersection QTL et eQTL*

Un eQTL et un QTL ont été considérés comme intersectant si le gène détecté comme *eGene* était présent dans le QTL. Par la suite, et plus précisément, les expressions des deux gènes (FBXO15 et C18orf63), présents dans l'intersection pour le QTL gras corporel, ont été ré-analysées par GWAS à l'aide de l'outil GCTA.

3.2.3. Résultats

Comparaison des outils de GWAS, GCTA et GEMMA, au travers les phénotypes gras corporel et poids de foie

Au vu des différences méthodologiques entre GEMMA et GCTA, leur capacité à détecter des QTL pour les deux phénotypes d'intérêt que sont le gras corporel et le poids de foie ont été testés. En considérant un seuil de significativité des p-value à 10^{-5}, GCTA a permis de détecter un QTL pour chacun des deux phénotypes étudiés (Figure 1A, gauche), alors qu'aucun n'a été détecté pour GEMMA (Figure 1A, droite). Pour le poids de gras corporel, le QTL identifié est sur le chromosome 2 et celui associé au poids du foie, sur le chromosome 5, les positions du

variant *leader* étant à 140620094 et 37854499 respectivement. Notons également un second QTL suggestif (p-value en limite du seuil) pour le gras corporel sur le chromosome 11 à la position 17386527. Les QQ-plots montrent un décrochage des p-values avec GCTA dû à la détection des QTL, ce qui n'est pas le cas avec GEMMA. Cependant, les valeurs de λ étant proche de 1, les modèles semblent dans les deux cas valides. Considérant l'ensemble de ces résultats, l'outil GCTA a été choisi pour procéder aux analyses GWAS ultérieures.

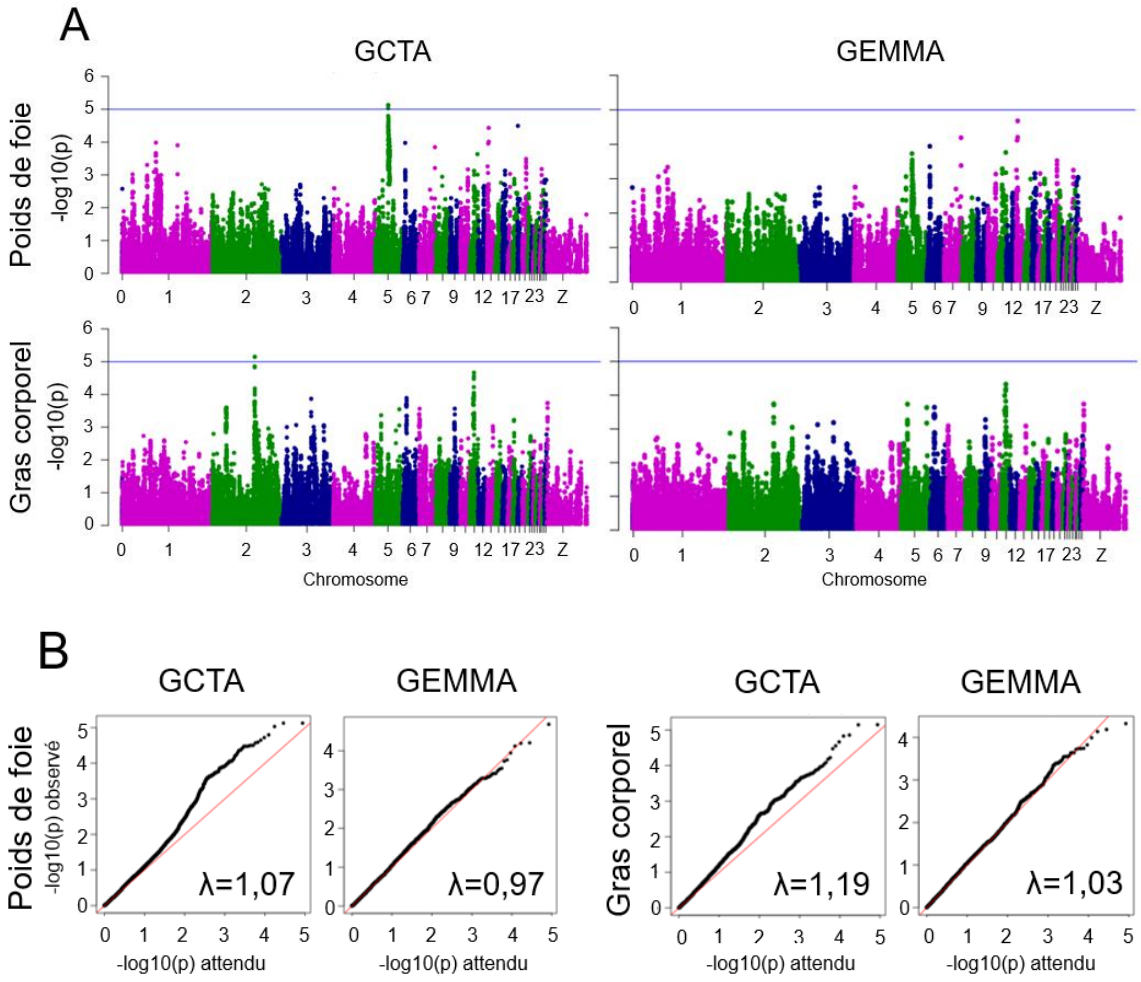


Figure 1 – Comparaison des résultats GWAS obtenus avec GEMMA et GCTA sur les phénotypes gras corporel et poids de foie. **A :** Manhattan plots, x : Chromosome ; y : -log10(p-value) **B :** QQ-plot, abscisse : -log10(p-value attendue sous la loi normale) ; ordonnée : -log10(p-value observée avec les données réelles). n = 940 individus de 90 semaines issus des lots de 2017, 2019 et 2020 avec 83 pères communs.

Caractérisation des QTL détectés par l'outil GCTA pour les phénotypes gras corporel, poids de foie et acides gras dans le foie

L'observation des variants inclus dans les QTL (+/- 500 kb autour du variant *leader*) détectés par GCTA pour les phénotypes gras corporel et poids de foie montre que les p-values les plus fortes correspondent aux valeurs de LD les plus élevés entre les variants et le variant *leader* (Figure 2, r2 ≥ 0,8). Cependant, pour le phénotype gras corporel, une baisse du déséquilibre de liaison est observée entre le variant *leader* et les SNP situés entre 91,3 et 91,5 Mb (Figure 2, Flèche en bleu), qui peut s'expliquer par une faible fréquence de l'allèle mineur (≤ 10%). Au sein de chaque QTL identifiée, d'une taille totale d'un Mb, 85 et 78 gènes sont présents respectivement dans le QTL associé au gras corporel et au poids de foie pour respectivement 453 et 467 SNP. Parmi les gènes présents dans ces QTL, 17 PCG sont identifiés pour le gras corporel et 24 pour le poids de foie. Parmi ces PCG, 2 sont associés au métabolisme des lipides (CYB5A et ZADH2) pour le QTL gras corporel et 4 (RDH11 ; ZFYVE26 ; LGALSL2 et EIF2S1) pour le poids de foie.

Les acides gras, mesurés pour 247 individus de 90 semaines du lot de 2020 avec 36 pères communs entre individus, ont été analysés par GWAS malgré l'absence d'une distribution gaussienne. Aucun QTL significatif n'a pu être détecté au seuil de significativité de 10^{-5} que ce soit pour les acides gras saturés (AGS), les acides gras mono-insaturés (AGMI), les acides gras polyinsaturés (AGPI) et les acides gras totaux (AG Totaux). De plus, aucun QTL suggestif n'est identifié.

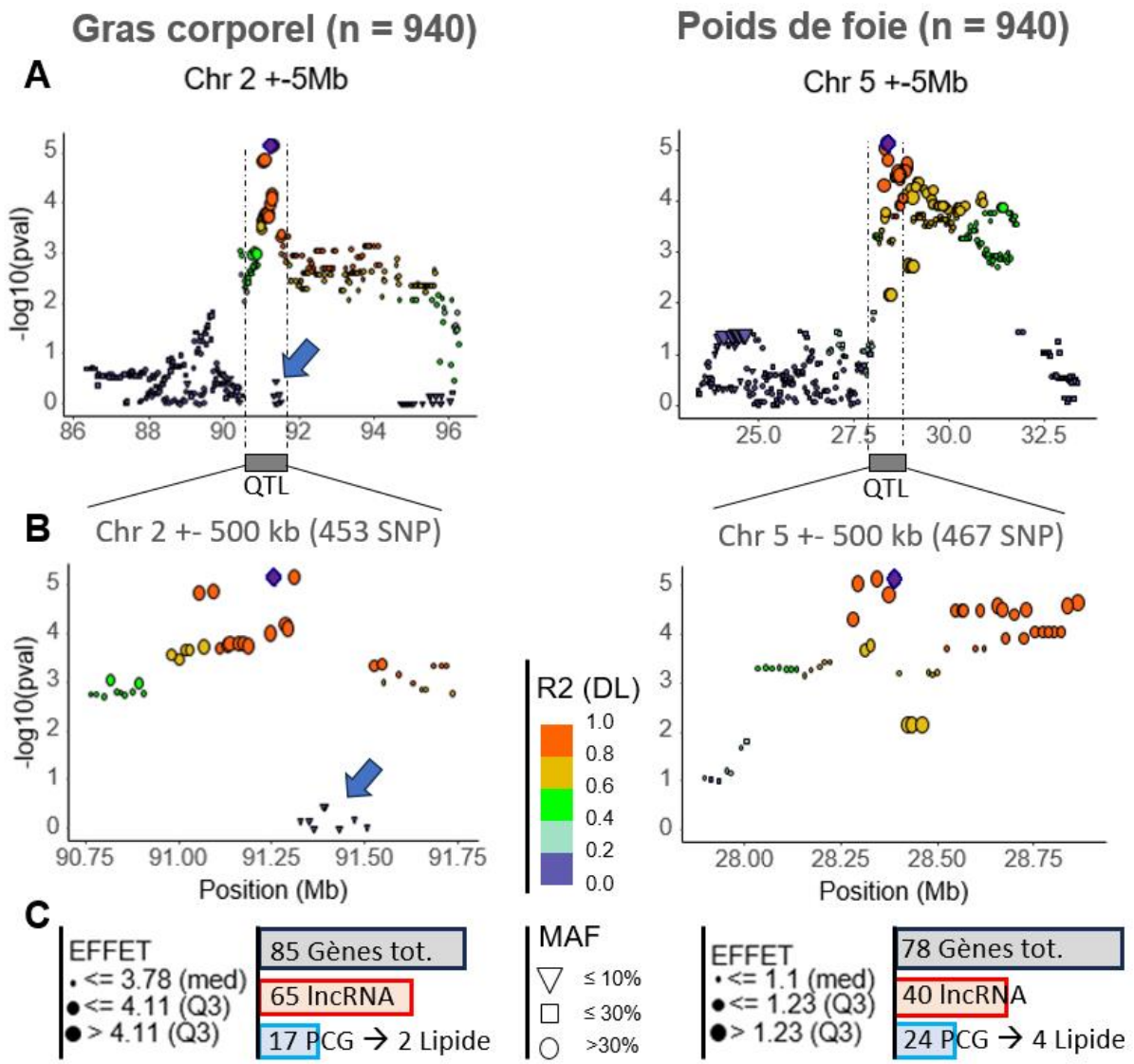


Figure 2 – Caractérisation des QTL associés au gras corporel (Chr 2) et au poids de foie (Chr 5) Un point représente un SNP à une position donnée. **A :** Manhattan plot autour du QTL (position du variant *leader* +- 5Mb) pour avoir une vision d'ensemble de l'environnement génomique autour du QTL. **B :** Zoom sur le QTL (± 500 kb autour du SNP *leader*) **C :** Quelques caractéristiques liées à chaque QTL, à savoir le nombre de gènes déclinés selon les biotypes PCG et lncRNA, et l'effet de chaque SNP sur le caractère. L'intensité de l'effet est donnée est par taille du point selon l'échelle fournie qui repose sur la distribution des effets (Q1, médiane, Q3) calculée à l'échelle du QTL. Losange violet : SNP *leader*. Le R^2 par rapport au SNP *leader* est indiqué par une échelle de couleur allant du violet (R^2 = 0 à 0,2) à l'orange (0,8 à 1). Abscisse : Position en Mb ; Ordonnée : -log10(p-value). n = 940 individus de 90 semaines des lots de 2017, 2019 et 2020 avec 83 pères communs.

GWAS sur l'expressions des gènes : cas des eQTL

Les GWAS sur phénotypes « expressionnels » ont été réalisés sur 363 individus de 90 semaines provenant de trois lots d'animaux différents pour lesquels le foie a été échantillonné. Parmi les 24 102 PCG et 44 428 lncRNA de l'atlas enrichi, 15 135 gènes sont exprimés dans le foie dont 12 590 PCG et 2 159 lncRNA. Sans surprise, les lncRNA sont significativement moins exprimés que les PCG (p-value = 3.1e-08) avec une expression médiane de 0,45 et 5,37 TPM respectivement.

Recherche des covariables impactant les phénotypes expressionnels

Considération des variables latentes sur les données d'expressions : Comme indiqué dans le *Mat. & Met*., l'identification des facteurs (covariables) ayant un effet sur les phénotypes expressionnels a été fait par une ACP faute de ne pouvoir estimer l'impact sur les 15 135 gènes pris indépendamment. En amont, une observation du premier plan factoriel, coloré en fonction des différents facteurs d'intérêt présentés en Table 1, a été réalisée (Figure 3). L'ACP montre alors que les lots d'élevage, conjointement aux lots de séquençage, sont impliqués dans la variabilité d'expression des gènes ayant permis de séparer les individus, notamment entre les individus de 2017 et ceux de 2019 et 2020. En effet, les individus nés en 2017 ont été élevés et abattus au printemps 2019 alors que ceux nés en 2019 et 2020 l'ont été en hiver 2021-2022 (décembre et avril). Ces résultats suggèrent donc que ces variations de saison lors de l'abattage peuvent influencer le transcriptome hépatique des animaux. Par ailleurs, les ARN des animaux du lot 2017 ont été extraits avec un protocole différent de ceux des individus des lots 2019 et 2020, ce qui pourrait également contribuer à un effet de ces lots sur le transcriptome hépatique. Cette hypothèse sera vérifiée dans un futur proche, car les ARN de cinq échantillons de foie de chaque lot de 2017 et 2019 ont été ré-extraits selon les deux protocoles et séquencés.

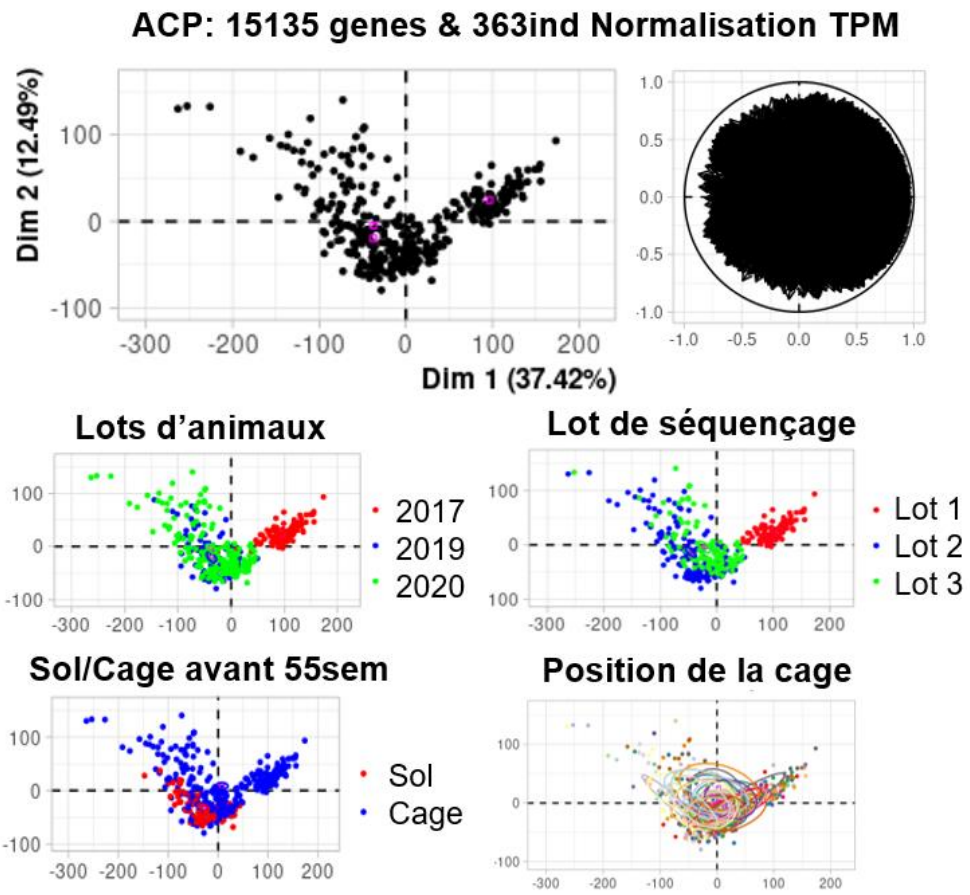


Figure 3 – ACP sur les expressions des gènes exprimés dans le foie. Expressions normalisées en TPM.

L'impact des facteurs latents ne pouvant être estimé pour chacun des 15 135 gènes pris indépendamment, les coordonnées de chaque individu sur les axes d'ACP ont été utilisés. Pour ce faire, les corrélations entre chaque covariable identifiée et les 15 premiers axes d'ACP sur les expressions ont été calculés (Table 4). Les facteurs liés aux lots d'élevage et au séquençage sont majoritairement expliqués par les cinq premiers axes d'ACP avec des corrélations allant de 0,63 à 0,12 pour la plus faible. Le seul axe parmi les 15 premiers expliquant le système d'élevage avant 55 semaines est l'axe 2 (corrélation = 0,14). L'axe 4, quant à lui, est corrélé à l'emplacement de la cage. Les cinq premiers axes, cumulant 63% de l'inertie totale, sont ceux majoritairement liés à l'ensemble des covariables testées. Ainsi, en considérant les coordonnés des axes 1 à 5 comme covariables dans le modèle impliquant les expressions géniques, une part des effets des variables latentes est prise en compte dans le modèle, tout en conservant un maximum de variabilité entre individus.

Table 4 – Corrélation entre axes d'ACP et covariables

Axes ACP	Inerties (%)	Lot d'animaux		Lot de séquençage		Sol/Cage avant 55sem		Emplacement cage		Effet père		Effet père (Génotype)		
		p-val	*cor*	*p-val*	*cor*	*p-val*	*cor*	*p-val*	*cor*	*p-val*	*cor*	Inerties (%)	*p-val*	*cor*
1	**37,42**	**4E-79**	**0.63**	**5E-89**	**0.68**	**1E-04**	**0.04**	**>0.05**		**4E-42**	**0.71**	**2,07**	**< 2.2e-16**	**0.87**
2	**12,49**	**4E-13**	**0.15**	**4E-13**	**0.15**	**2E-13**	**0.14**	**>0.05**		**1E-03**	**0.32**	**2,04**	**< 2.2e-16**	**0.89**
3	**7,16**	**2E-10**	**0.12**	**3E-13**	**0.15**	**2E-02**	**0.01**	**>0.05**		**3E-06**	**0.38**	**1,80**	**< 2.2e-16**	**0.85**
4	**3,72**	**>0.05**		**4E-03**	**0.03**	**>0.05**		**4E-04**	**0.17**	**2E-03**	**0.32**	**1,66**	**< 2.2e-16**	**0.82**
5	**2,64**	**2E-13**	**0.15**	**2E-20**	**0.22**	**>0.05**		**>0.05**		**2E-02**	**0.29**	**1,60**	**< 2.2e-16**	**0.82**
6	2,09	>0.05		>0.05		>0.05		>0.05		2E-02	0.29			
7	1,95	>0.05		4E-04	0.04	>0.05		>0.05		>0.05				
8	1,24	3E-03	0.03	4E-02	0.02	>0.05		5E-02	0.12	4E-02	0.28			
9	1,08	>0.05		>0.05		>0.05		>0.05		6E-04	0.33			
10	0,97	>0.05		>0.05		>0.05		>0.05		>0.05				
11	0,88	>0.05		>0.05		>0.05		>0.05		8E-03	0.30			
12	0,79	3E-03	0.03	3E-09	0.10	>0.05		>0.05		1E-02	0.30			
13	0,71	7E-09	0.10	4E-11	0.12	2E-03	0.03	1E-02	0.14	3E-04	0.34			
14	0,60	>0.05		>0.05		>0.05		>0.05		>0.05				
15	0,55	>0.05		>0.05		>0.05		2E-02	0.13	1E-03	0.32			

Prise en compte de l'effet père : TensorQTL ne prenant pas de matrice de parenté en entrée, les coordonnées des individus sur les axes d'ACP, réalisée avec les variables de type génotype, ont été utilisés afin de capter l'apparentement entre individus (effet père). L'effet père, présent dans notre cohorte, est bien capté par les cinq premiers axes d'ACP sur les expressions, la majorité des axes sont significatifs et sont corrélés à hauteur d'environ 30%. L'ACP réalisé sur les génotypes des individus montre des corrélations entre les effets pères et les axes d'approximativement 85% pour l'ensemble des axes (Table 4). L'effet père étant responsable d'une part de la variabilité génétique et afin de minimiser cet effet latent, la littérature préconise de prendre en compte le nombre d'ACP génotype en fonction du nombre d'individus [389]. Le nombre d'individus étudiés étant ici de 363, cinq axes d'ACP sur génotype sont recommandés.

Finalement, ce sont donc les cinq premiers des ACP sur les expressions de gènes ainsi que les cinq premiers axes d'ACP sur les génotypes qui ont été considérés comme covariables dans les modèles mis en place dans la détection de eQTL.

Nombres et caractérisation des eQTL détectés

Après l'application de TensorQTL et la prise en compte des covariables, 242 eQTL distants et 4 185 locaux ont été détectés pour nos 363 individus. Concernant les 242 eQTL distants, 89 ne sont pas positionnés dans la dernière version du génome GRCg7b, d'où l'assignation à un chromosome nommé « 0 », et ont été écartés.

Parmi les 153 eQTL restants, on distingue deux types de eQTL distants, ceux présents sur le même chromosome que le eGene et ceux qui sont sur un chromosome différent. Parmi les 99 eQTL présents sur le même chromosome que le gène qu'ils régulent, 77 régulent des PCG et 21 des lncRNA ce qui est semblable aux ratios de gènes exprimés dans le foie (83% et 14%). Deux fois moins de eQTL sont détectés sur des chromosomes différents, soit 54 eQTL, 49 étant pour des PCG. De façon intéressante, parmi les 54 eQTL distants situés sur un chromosome différent de leur eGene, 13 correspondent au même SNP *leader* qui indique donc une région eQTL qui régulent 13 gènes PCG différents dont 11 ont un nom HGNC connu. Dans cette région eQTL, un PCG connu est présent (GAB1) mais des analyses complémentaires sont nécessaires pour appuyer cette observation.

4 185 eQTL locaux ont été détectés parmi les 15 135 gènes exprimés, à savoir 28% des gènes analysés. À tire d'information, notons que l'application d'une correction de type Benjamini-Hochberg en remplacement de Bonferroni permet d'augmenter le nombre de eQTL locaux identifiés à 8 062. Brièvement, parmi les 4 185 eQTL détectés, 3 441 et 673 sont associés à la régulation de PCG et des lncRNA respectivement, soit des ratios semblables à ceux des gènes exprimés dans le foie. 431 sont détectés comme associés à la régulation de gènes liés aux métabolismes énergétique et lipidique.

3.2.4. Discussion et conclusion

Comme discuté dans les Résultats §2.3, la puce utilisée dans ce travail a été conçue selon l'assemblage galgal5, seuls 33 chromosomes autosomaux étaient alors identifiés. Afin de couvrir l'intégralité du génome et d'augmenter le nombre de SNP génotypés, la détection de variants sur GRCg7b à partir des données RNAseq qui ont été utilisées pour les analyses eQTL est en cours. Cette analyse est réalisée à l'aide d'une adaptation du *pipeline* « nf-core/rnavar » [452, 541]. Grâce à cela, nous devrions non seulement couvrir de nouvelles zones mais également densifier par 10 le nombre de SNP étudiés. Dans le cadre de la détection de QTL par GWAS, un variant était considéré comme significatif si la p-value était inférieur à 10^{-5} ce qui est couramment dans ce type d'analyse [595]. Cependant, ce seuil reste arbitraire et est très dépendant du plan expérimental, notamment du nombre de marqueurs et du nombre d'individus étudiés. Ainsi, s'il est classique de prioriser les zones du génome ayant une p-value dépassant ce seuil, certaines zones ne dépassant pas ce seuil, mais montrant tout de même un décrochage (nommées QTL suggestifs dans les résultats) ne sont pas à omettre et pourront être analysées dans un second temps. Par exemple, pour les acides gras, aucun QTL n'a été identifié du fait potentiellement d'un nombre faible d'individus (< 250 issus seulement de 36 pères) et de l'application de l'approche GWAS sur des données présentant une bimodalité (stéatosé vs. sain).

Toutefois, certains loci géniques semblent tout de même ressortir, notamment au niveau du chromosome 15, et ce, pour les quatre phénotypes liés aux acides gras. Cependant, des analyses plus approfondies impliquant un ajustement du modèle prenant en compte la distribution bimodale combiné à la prise en compte d'échantillons supplémentaires sont nécessaires pour conforter l'existence de ce QTL. Concernant la bimodalité de ce phénotype, l'analyse des AG totaux hépatiques montre que 18,6 % des 247 individus analysés présentent un foie surchargé en lipides avec en moyenne, 44 % de gras dans le foie contre 2,6 % pour les autres individus. Notons que chez l'humain, un foie est considéré stéatosé si le taux de triglycérides intra-hépatique est d'au moins 5 % du poids du foie [596].

Les analyses portant sur la détection de eQTL ont permis d'identifier, parmi les 15 135 gènes exprimés, 4 185 *local*-eQTL (28 %) et 242 distant-eQTL (1,6 %) dont 54 sur des chromosomes différents. Ces résultats sont dans les mêmes ordres de grandeur que ceux retrouvés dans le cadre des consortiums GTEx [389, 405, 406, 574]. En effet, pour un nombre d'échantillons

avoisinant les 250, le nombre de *local*-eGene est compris entre 4 000 et 6 000 alors qu'il est d'une vingtaine pour les *distant*-eQTL. De façon intéressante un eQTL trans dit « hub » ou « hotspot » et régulant dans notre cas 13 gènes semble avoir été détecté. Cependant, ces gènes ne semblent pas être reliés par une fonction commune. La région doit être plus finement analysée afin de mieux définir la zone eQTL en LD avec le variant *leader* et ainsi définir les gènes candidats pouvant réguler le groupe de eGene identifiés. Ce type de eQTL sont rarement reportés dans la littérature et nécessite d'être confirmé par des expériences de biologie moléculaire et cellulaire.

Pour finir, nous avons analysé en profondeur le QTL détecté pour le gras corporel afin d'identifier des gènes candidats causaux pour ce caractère. En considérant l'hypothèse d'un variant régulateur impactant l'expression d'un gène et que cette régulation ait lieu dans le foie, organe majeur du métabolisme des lipides, nous avons réduit la liste des 85 gènes à 15 gènes exprimés dans le foie, puis à 2 eGene régulés par un eQTL *local,* a priori co-localisant avec la région. Des analyses supplémentaires sont nécessaires pour confirmer la colocalisation des eQTL et QTL.

Discussion et perspectives

1. La modélisation des modèles géniques et leurs annotations fonctionnelles nécessitent un effort collectif avec des données nombreuses et standardisées

1.1. Couvrir la diversité des modèles géniques par une réelle utilisation du *big data*

Nos travaux portant sur l'atlas enrichi en gènes du génome de la poule selon le dernier assemblage GRCg7b (voir Résultats §1.1) ainsi que sa mise à disposition via GEGA, un outil en ligne facile d'accès (voir Résultats §1.2), ont permis de compléter l'annotation du génome de cette espèce d'élevage, en particulier en modèles géniques de type lncRNA. Comme observé dans ces deux études, il existe des différences importantes pour les modèles de gènes et de transcrits entre les deux principales bases de données de référence Ensembl et RefSeq. D'autre part, certains projets indépendants produisent également leur propre annotation qui ne recouvre qu'en partie ces annotations de référence [112, 323, 372]. Dans ces deux situations, l'ensemble des acteurs utilisent bien souvent des *pipelines* différents et donc non standardisés. Plus spécifiquement pour Ensembl et RefSeq, les *pipelines* utilisés sont difficilement accessibles, voire mal documentés et utilisent des logiciels variés. Ensembl par exemple aligne les données RNAseq avec « BWA » couplé à « Exonerate » [597] alors que RefSeq utilise, pour sa part, l'aligneur « STAR » [598, 599]. De plus, ces *pipelines* utilisent bien souvent des données complémentaires telles que les séquences protéiques connues, des données CAGE ou même des annotations manuelles pour préciser les modèles géniques. Selon les principes de plus en plus prônés d'*open science* [600], l'élaboration d'un *pipeline* de modélisation des modèles géniques, « *gold standard* », documenté et reproductible, serait fortement souhaitable. Ce dernier pourrait par exemple être intégré à des projets communautaires tels que *nf-core* [452] où la communauté pourrait participer à son amélioration. À notre connaissance, un *pipeline* répondant à ces problématiques a été initié en 2022. Il se nomme « genomeannotator » [601] et est consultable sur *nf-core* mais aucune version stable n'a été annoncé et peu d'informations sont disponibles.

D'autre part, pour un organisme donné, et que ce soit Ensembl, RefSeq ou des projets d'annotation indépendants, ils tendent tous à utiliser un ensemble de données (en général de type RNAseq) spécifique et/ou qui évolue peu et ainsi qui ne couvre pas la diversité des tissus

et modèles cellulaires, pourtant disponibles dans les bases publiques telles que ENA/SRA. En effet, l'utilisation des données RNAseq est souvent réduite à « quelques » échantillons représentant peu de tissus. Par exemple, chez la poule, 63 et ~300 échantillons sont respectivement utilisés par Ensembl v107 [602] et NCBI v106 [603] ne représentant ainsi qu'une faible partie des données disponibles sur ENA/SRA. En effet, en utilisant la requête « *tax_eq(9031) AND library_strategy="RNA-Seq"* » sur ENA, 24 839 données RNAseq ont pu être identifiées pour la poule. Pour tenter de s'approcher de l'exhaustivité concernant les modèles géniques, l'idéal serait d'inclure tous ces échantillons provenant de nombreux tissus (et/ou de données *single cell*) à différents stades de développement et pour les deux sexes, permettant ainsi de capter les gènes exprimés dans toutes ces conditions. Ces données serviraient également pour l'annotation expressionnelle des gènes que nous aborderons dans la seconde partie (voir Discussion §1.2). Si cette utopie ne s'avère en aucun cas réalisable pour des raisons de temps de calculs et de coûts, plusieurs études telles que les projets GTEx [389, 405, 406, 574] ont déjà analysé une quantité importante de données RNAseq déjà disponibles, voire ont généré spécifiquement des données (cas du GTEx humain), pouvant d'ores et déjà être utilisées pour modéliser les gènes. Dans la V8 du GTEx humain [389], un total de 17 382 échantillons transcriptomiques pour 54 tissus issus de 948 donneurs ont été spécifiquement générés et sont donc disponibles ; seuls 5 tissus comptent moins de 70 échantillons. De plus, les informations d'âge, de sexe et d'ethnie sont rigoureusement renseignées pour ces 948 donneurs. Notons aussi le lancement du dGTEX (*developmental* GTEX) [604] en septembre 2021 qui a pour objectif d'analyser l'expression des gènes au cours du développement chez l'humain et donc de collecter des données RNAseq pour 4 stades de développement (0 à 2 ans, 2 à 8 ans, 8 à 12,5 ans et 12,5 ans à 18 ans). La collecte des données annoncée pour septembre 2023 vise à obtenir, pour chacune des 4 catégories, une 30aine de tissus pour environ 120 donneurs (soit un total d'environ 14 400 échantillons). Concernant les espèces d'élevage, le FarmGTEX [457] affiche lui aussi les mêmes ambitions pour six espèces d'élevage avec la volonté de collecter/rassembler des données pour une 50aine de tissus (et/ou types cellulaires) pour plus de 1000 individus. Plus précisément pour la poule, le ChickenGTEX (voir Résultats §3.1) [405] compte 52 tissus pour un total de 7 015 échantillons RNAseq collectés parmi les 24 839 RNAseq disponibles dans ENA/SRA.

Cependant, un effort collectif doit être fait de la part des laboratoires déposant leurs données sur ces plateformes pour fournir de manière précise et correcte les différentes métadonnées

associées aux échantillons afin qu'ils puissent être intégrés de façon automatisée à la modélisation des modèles géniques. À titre d'exemple, sur les 24 839 RNAseq de poule déposés sur ENA, 10 792 (43 %) ne présentent pas l'information tissulaire dans le bon champ. L'information peut en effet être disponible dans d'autres métadonnées, tels que dans le nom de l'échantillon ou dans la méta-description (même si ce n'est pas systématique) mais il n'est ainsi pas possible d'utiliser l'information de manière automatisée. Notons également que la nomenclature utilisée n'est pas standardisée nécessitant donc un effort sur les ontologies. Par exemple, le foie est identifié au moins sous neuf labels différents. Face à ces lacunes, le consortium FAANG [195, 196], spécifique des espèces d'élevage, impose pour le dépôt des données sur ENA de renseigner des métadonnées (*e.g.*, le sexe et le tissu) selon une nomenclature définie, évitant par conséquent une perte d'information ou de l'ambiguïté. (*e.g.*, « *liver* » = UBERON_0002107 selon l'*Ontology Lookup Service* [605]). Parce que certaines données sont dites sensibles, car appartenant à des sociétés privées comme les sélectionneurs, le consortium FAANG a distingué deux types de métadonnées : *i)* les métadonnées obligatoires comme le sexe et *ii)* les métadonnées recommandées ou optionnelles comme la race/lignée et le pedigree. Le choix des métadonnées obligatoires est également basé sur leurs forts impacts sur l'expression des gènes. Comme montré avec nos travaux sur le ChickenGTEX (voir Résultats §3.1) [405] mais également sur l'atlas enrichi (voir Résultats §1.1) [125], les expressions des gènes diffèrent davantage entre tissus ou sexe ou stades de développement qu'entre races ou conditions.

Finalement, en supposant que les métadonnées soient correctement indiquées et qu'un *pipeline* de référence soit mis en place, il serait envisageable de produire une unique annotation des modèles géniques par espèce en utilisant l'ensemble des données RNAseq disponibles sur les bases de données publiques. Cette idée suppose par ailleurs un travail conjoint des différents acteurs. Cela n'apparait pas impossible, le projet MANE [429], pour le moment uniquement chez l'humain et visant à associer pour chaque PCG un transcrit représentatif commun à Ensembl et à RefSeq en est un bel exemple. Cependant, cette approche pose deux problèmes. Tout d'abord, elle n'est pas automatisable facilement pour les autres espèces, car un gros travail d'annotation et de vérification manuelle a été mis en place. De plus, le fait de n'associer qu'un transcrit par PCG est très restrictif vu qu'il est connu qu'un gène est associé à différents transcrits exprimés selon différents contextes. En réalité,

cette solution apparait comme un « pansement », la réalisation d'une annotation commune dès le départ, *i.e.* avec les mêmes données RNAseq, permettrait de s'affranchir de ce travail.

Il faut cependant avoir conscience que la modélisation des gènes en intégrant une si grande quantité d'information est très consommatrice de ressources et qu'il n'est ainsi pas envisageable de relancer les analyses à chaque ajout de données sur les bases publiques. Pour pallier cela, deux solutions apparaissent. Premièrement, fixer un délai où le *pipeline* serait relancé avec l'ensemble des données disponibles, par exemple, tous les ans, à date précise. Les laboratoires ayant généré des données connaîtraient alors cette date et auraient la possibilité de déposer en amont leurs données pour que celles-ci soient intégrées dans l'annotation finale. L'autre solution, moins pertinente que la précédente, mais également moins coûteuse en ressources, serait de proposer un outil permettant d'agréger des modèles géniques sur une annotation existante. Cela nécessite alors *i)* d'une part, de modéliser les modèles géniques uniquement à partir de nouveaux échantillons, même si cela peut conduire à de mauvais modèles géniques (*e.g.,* fusion ou morcèlement de modèles) si les données ne présentent pas une profondeur suffisante [124, 125, 160] ; *ii)* d'autre part, d'intégrer les modèles ainsi générés à l'annotation globale.

Cette intégration est envisageable selon deux manières. Le première est celle que nous avons utilisée dans le cadre de l'atlas (voir Résultats §3.1) [125]. L'approche consiste alors à ajouter successivement les sources en considérant uniquement les modèles géniques. Ainsi, si un gène d'une source à ajouter possède un transcrit qui superpose, ne serait-ce que d'une base, un exon d'un transcrit d'un gène déjà présent, l'entièreté du modèle génique de la source à ajouter est évincée. En réalité, cette approche apparaît quelque peu basique car un gène se verra toujours associé des transcrits d'une seule source, aucun modèle de transcrit n'étant ajouté au modèle génique de base. La seconde approche, plus informative, mais que nous n'avons pas mis en œuvre faute de temps, est d'intégrer les modèles à l'échelle des transcrits en considérant les exons et les chaînes introniques des deux ressources à fusionner. Cela permet alors d'enrichir les modèles de gènes avec des transcrits alternatifs même si cela demande parfois de redéfinir le modèle du gène à ses extrémités.

La solution constituant à intégrer de nouveaux modèles à l'annotation globale est par ailleurs déjà proposée par des projets tels que TAGADA [606] qui se présente comme un *pipeline* d'analyse de données RNAseq sous NextFlow [607] pour l'assemblage et l'analyse de transcrits

et de gènes. À partir de la séquence génomique de référence (*.fasta*), de l'annotation de référence de ce génome (*.gtf*) et de données RNAseq obtenues dans un projet spécifique (*.fastq*/*.bam*), TAGADA peut notamment améliorer l'annotation de référence en trouvant de nouveaux gènes et transcrits avec une annotation de leur biotype, notamment pour les lncRNA. En revanche, cette annotation générée sera spécifique à chaque projet apportant les données de RNAseq.

Pour finir, il faut garder en tête que l'annotation des gènes du génome est relative à un assemblage et que celui-ci s'améliore en lien avec l'évolution des technologies. Soulignons, par exemple, la parution des génomes dits T2T (« *Telomere-to-Telomere* ») qui, en se basant sur l'utilisation du *long-read,* peuvent préciser des régions du génome peu ou pas couvertes tels que les télomères, les centromères ou des régions chromosomiques contenant une part importante de séquences répétées [608]. Si le génome T2T est paru la première fois en 2021 pour l'humain [474], Huang et al. [609], ont présenté en 2023 un génome T2T complet de poule (à l'exception du chromosome W) avec tous les chromosomes assemblés et toutes les lacunes comblées, notamment pour les micro-chromosomes et le chromosome 16. Cette progression de l'assemblage doit alors conduire à une révision complète de l'annotation de référence quelle que soit la stratégie adoptée. Ainsi, durant ma thèse, le passage de l'assemblage GRCg6a à GRCg7b par Ensembl et RefSeq en juillet 2022 m'a conduit à revoir entièrement l'atlas enrichi en lncRNA développé en 2020 par le laboratoire d'accueil.

En conclusion, l'annotation des modèles géniques est un processus complexe et continu qui nécessite des efforts coordonnés entre différents acteurs. L'utilisation de *pipelines* standards, le renseignement et le partage rigoureux des métadonnées et la consolidation des données issues de multiples sources sont des éléments clés pour générer des modèles de gènes de référence les plus complets et précis possibles. Cette annotation exhaustive est cruciale pour la recherche en génomique, mais doit aussi évoluer en tandem avec l'amélioration des assemblages de génomes. Ainsi, à ce jour, et compte tenu des avancées concernant le génome de la poule, l'idéal serait que Ensembl et RefSeq profite du nouvel assemblage T2T pour produire conjointement cette annotation exhaustive. Faute de s'accorder, chaque instance pourrait cependant intégrer des données supplémentaires afin d'améliorer l'annotation. Dans

tous les cas, il est souhaitable qu'une annotation commune à ces deux références soit générée, notamment en agrégeant les différents modèles géniques à l'échelle des transcrits.

1.2. Considérer les gènes dans un contexte biologique : apport d'informations fonctionnelles par les profils d'expression

Comme vu dans le paragraphe précédent, les travaux exposés dans la première partie des résultats ont enrichi le nombre de modèles géniques identifiés chez la poule pour l'assemblage GRCg7b, mais ils ont également apporté des informations fonctionnelles via notamment les profils d'expressions. En effet, afin de comprendre le fonctionnement de ces gènes et de les mettre potentiellement en parallèle avec des caractères d'intérêt, il convient d'identifier les conditions dans lesquelles ils s'expriment, c'est-à-dire a minima les tissus, mais aussi, si possible, le sexe, les stades de développement, voire les conditions de milieu auxquelles sont soumis les individus desquels les tissus ont été extraits. Des atlas d'expressions couvrant cette diversité sont alors nécessaires et c'est ce que proposent les projets GTEX [389, 405, 406, 574] en utilisant la multitude d'informations à leur disposition. D'autres ressources, plus anciennes, telles que l'*Expression Atlas* [610] *ou GEO* (*Gene Expression Omnibus*) [611], fournissent également des informations sur l'expression des gènes au travers de différents tissus, types cellulaires, stades de développement ou maladies, entre autres. Cependant, cela s'avère plus problématique dans le cas des bases d'annotation des génomes de référence telles que Ensembl et RefSeq. En effet, ces dernières ne produisent pas leurs propres données d'expression, mais les extraient respectivement de l'*Expression Atlas* et de *GEO*. Ainsi, plusieurs problèmes apparaissent : *i)* Si un changement d'identifiants de modèles géniques est observé du fait de l'ajout/suppression d'un gène ou du fait de l'évolution de l'assemblage, la relation est brisée. C'est notamment le cas pour la poule où actuellement (octobre 2023), aucune information pour aucun gène n'est disponible dans Ensembl concernant les profils d'expression, du fait de la sortie du nouvel assemblage GRCg7b et des identifiants associés qui sont dans leur totalité différents de ceux actuellement utilisés dans l'*Expression Atlas* [612]. De plus, *ii)* ces bases n'utilisent que quelques projets publics et ne couvrent donc ni la diversité des tissus disponibles dans les bases de données publiques (*e.g.*, les 24 839 RNAseq disponibles sur ENA) ni la diversité de tissus (certes moindre) utilisés par les bases

d'annotation de référence. Il est ainsi possible d'observer des cas, où le gène est modélisé, mais où il n'est jamais exprimé, ce qui est paradoxal puisque des données RNAseq, se basant par définition sur l'expression, ont été en grande partie utilisées pour la modélisation des gènes. À titre d'exemple, pour l'*Expression Atlas* et la poule [612], seul 4 projets avec des conditions normales et 35 projets avec des traitements différentiels sont utilisés. Parmi ces 39 projets, 19 sont issus de données RNAseq, les 20 autres proviennent de projets antérieurs à 2015 qui utilisent des puces, ainsi seuls une fraction des gènes y sont quantifiés et la comparaison avec les données plus récentes de RNAseq n'est pas aisée.

Afin d'apporter de l'information « expressionnelle » pour l'ensemble des gènes d'une annotation, il serait alors envisageable d'intégrer une étape de quantification de l'expression des gènes au *pipeline* d'analyse visant à modéliser les gènes (discuté dans la partie précédente). En effet, il serait intéressant que l'ensemble des données RNAseq qui a permis la modélisation des gènes soit également utilisé pour quantifier l'expression des gènes. En supposant que les métadonnées des projets utilisées soient bien renseignées, il serait alors possible d'avoir un réel atlas d'expression des gènes couvrant un large éventail de tissus, pour un nombre conséquent d'individus (potentiellement avec des différentiels de conditions et de sexes) et en lien avec une unique annotation. Ces informations constitueraient dès lors une première porte d'entrée pour étudier les expressions géniques, les tissus-spécificités ou encore les co-expressions et permettraient d'émettre des hypothèses quant à la fonction de certains gènes, comme pour les lncRNA par exemple. En attendant de telles solutions, nos travaux (voir Résultats §3.1 et §3.2) ont permis de dresser les profils d'expressions pour l'ensemble des gènes de Ensembl et de RefSeq (ainsi que pour des gènes issus de ressources complémentaires) au travers de 47 tissus représentant les grandes fonctions physiologiques chez la poule. 1400 échantillons issus de 36 projets ont ainsi été utilisés pour prendre en compte d'éventuelles variations liées au sexe, à la race ou à des conditions expérimentales.

1.3. L'annotation des ARN longs non-codant nécessitent de nouvelles approches

Les principales difficultés concernant la connaissance des lncRNA sont, d'une part, la confirmation de leur existence du fait de leur faible expression et, d'autre part, la compréhension de leurs fonctions au sein d'une espèce. En effet, à ce jour, et comme souligné dans la *review-1* pour les espèces d'élevage (voir Introduction §1.2) [103], les lncRNA sont en général 10 fois moins exprimés que les PCG et très peu d'entre eux ont une fonction caractérisée.

Le concept de « *guilt-by-association* » [445], qui stipule que les gènes co-exprimés sont plus susceptibles de partager une fonction, a alors été appliqué aux lncRNA pour tenter d'inférer leurs fonctions. La méthode la plus couramment utilisée consiste ainsi à identifier de manière globale les gènes dont l'expression est corrélée aux lncRNA d'intérêt [331, 491, 613–615]. Les gènes PCG sont étudiés en priorité, car ce sont les mieux modélisés et les mieux annotés fonctionnellement. L'analyse des profils d'expression des lncRNA et des PCG à l'échelle du génome entier permet ainsi de faire des groupes caractéristiques et les fonctions des lncRNA d'un groupe peuvent alors être prédites sur la base des lncRNA et surtout des PCG ayant des fonctions connues du même groupe. Cependant, les PCG d'un groupe de co-expression peuvent être associés à différentes fonctions. Aussi, la stratégie utilisée pour synthétiser les fonctions portées par chaque groupe est l'analyse d'enrichissement de termes fonctionnels standardisés de type GO et KEGG [616].

Considérant que les lncRNA sont des gènes dits régulateurs de l'expression, il est communément admis que parmi les gènes co-exprimés d'un groupe, il y a des PCG, gènes cibles de l'action de lncRNA. Cependant, pour analyser ces relations de cause à effets, il convient d'avoir recours à des expérimentations de biologie moléculaire et cellulaire. En général, un PCG est considéré comme régulé par un lncRNA s'il est exprimé de manière différentielle après l'extinction (totale ou partielle) ou la surexpression du lncRNA d'intérêt [617]. Cependant, le nombre d'expériences de ce type est encore limité pour les lncRNA, car leurs mises en place sont très couteuses et nécessitent beaucoup de main d'œuvre.

Cependant, n'oublions pas que les lncRNA interagissent également avec les miRNA [170, 618], ainsi même si la quantification de leur expression nécessite des ressources supplémentaires, la co-expression avec ces derniers est une voie à considérer.

Les lncRNA sont connus pour réguler un gène PCG à distance, mais également en *local*, en interagissant avec les régulateurs transcriptionnels spécifiques du PCG situé à proximité [165]. Il est donc possible d'inférer également un potentiel rôle de régulateur du lncRNA sur son PCG à proximité en identifiant les paires lncRNA:PCG co-exprimés et co-localisés. Cette approche est supportée par nos travaux sur l'atlas enrichi (voir Résultats §1.1) qui montrent que des gènes proches, et d'autant plus pour les couples lncRNA:PCG que les couples PCG:PCG, ont tendance à être plus fortement co-exprimés [108, 124, 125]. À l'inverse, si un lncRNA est co-exprimé avec un PCG, les deux gènes sont fréquemment séparés par une distance inférieure à 10 kb dans le génome linéaire. Les études du GTEx [389, 619] viennent aussi appuyer cela en indiquant que 1/4 des variants régulateurs agissant sur plusieurs gènes régulent à la fois des PCG et des lncRNA proches et que ces derniers étaient situés dans le même TAD. Cependant, cette colocalisation interroge également sur la position du lncRNA par rapport au PCG associé. En effet, alors que les gènes en position « *divergent* » et « *same-strand* » semblent voir leur co-expression augmenter lors de la réduction de distance, cela n'est pas le cas pour les gènes convergents, signe potentiel de mécanismes d'actions variables [620–622]. Pour finir, notons que notre approche couplant position et expression est contrainte par les modélisations encore incertaines des lncRNA. En effet, comme vu dans l'atlas enrichi, certains lncRNA notamment « *same-strand* » ne sont que des artefacts et s'avèrent n'être en réalité qu'une prolongation d'un PCG proche ou même d'un autre lncRNA avec lesquels ils sont par définition co-exprimés. Cependant, certains couples PCG:lncRNA sur un même brin sont de véritables loci indépendants comme démontrées par nos validations expérimentalement. Cette vérification nécessite des expérimentations de biologie moléculaire apparait néanmoins couteuse et difficile à mettre en œuvre à grande échelle. Une solution intermédiaire pour valider ces lncRNA:PCG proche et sur un même brin serait alors de quantifier de manière systématique le nombre de *reads* de RNAseq chevauchant chacun des deux modèles, un nombre élevé de reads étant un indicateur de mauvaise modélisation.

Si considérer les lncRNA par rapport à un PCG apparait être une voie à utiliser pour inférer des fonctions, il convient de développer des approches alternatives en considérant les lncRNA

indépendamment. Une approche consiste alors à tenter d'apporter des éléments sur leurs fonctionnalités par la seule analyse de leur séquence. Ce type d'approches a déjà fait ses preuves pour les PCG à fonction inconnue dans une espèce donnée. Il s'agit par exemple de rechercher dans la protéine codée par le gène d'intérêt des motifs d'acides aminés, signature d'une fonction [623, 624]. Cependant, ces outils ne peuvent être utilisés directement pour les lncRNA, car ils ne présentent qu'une faible conservation de leur séquence primaire entre espèces, et ce, d'autant plus qu'elles sont éloignées phylogénétiquement. Cependant, au sein de cette faible conservation semble tout de même se cacher des petits motifs de séquences apparentées (les k-mers) qui seraient conservés, car essentiel au maintien des fonctions des lncRNA [487]. Des outils comme SEEKR permettent alors de regrouper les lncRNA à partir de leur séquence en k-mers et donc d'associer une fonction à un ou des profils de k-mers sur la base de lncRNA (et pourquoi pas de PCG, si on en croit l'histoire évolutive) à fonction connue. Un autre outil nommé lncLOOM [286], également basé sur les k-mers cherche à utiliser les informations entre plusieurs espèces afin de détecter en toute confiance des éléments fonctionnels spécifiques qui ont été conservés au cours de l'évolution dans des lncRNA orthologues d'espèces éloignées.

Comme nous venons de le dire, cette approche suppose donc d'établir de manière systématique des relations d'orthologie entre lncRNA de différentes espèces. En effet, l'intérêt de cette approche est multiple, car elle permet *i)* de détecter des lncRNA conservés par l'évolution entre espèces, signe d'un intérêt biologique important, *ii)* d'inférer une fonction à un gène d'une espèce grâce à la fonction connue de ce même gène dans une espèce plus étudiée et *iii)* dans le cadre de méthode d'alignement, telle que présentée en Résultats §1.4, de détecter des lncRNA qui n'aurait pas été modélisé dans une espèce par manque d'échantillons biologiques. Cette méthode est par exemple grandement employée pour les PCG de la poule. En effet, parmi les 17 007 PCG identifiés chez Ensembl V110, 74,4 % présentent un orthologue unique avec l'humain dans BioMart et pour lesquels une grande partie des fonctions a ainsi été inférée.

Cependant, à ce jour, aucune relation d'orthologie n'est disponible pour les lncRNA dans les bases de données de référence, et ce, que ce soit pour des espèces distantes phylogénétiquement telles que la poule et l'humain ou pour des espèces proches et avec un effort d'annotation plus important tels que la souris et l'humain. Les travaux exposés sur

l'orthologie des lncRNA (voir Résultats §1.4) ont alors pour objectif de fournir, dans un premier temps, des listes de lncRNA potentiellement orthologues pour permettre, dans un second temps, un potentiel transfert de connaissance. À défaut de pouvoir réellement assigner une fonction, les lncRNA supposés comme orthologues pourront être sujets à des analyses d'expressions ou de co-expressions comme celles évoquées en début de partie. En effet, même si les séquences et les niveaux d'expressions évoluent plus rapidement pour les lncRNA, il a été démontré que la spécificité tissulaire et les profils d'expression des gènes fonctionnels sont généralement conservés d'une espèce à l'autre [276].

En conclusion, déchiffrer les fonctions des lncRNA demeurent un défi majeur. Bien que l'analyse de la conservation des séquences soit une voie prometteuse, il est également important de considérer les lncRNA dans leur contexte génomique et leur co-expression dont celles avec les gènes voisins. Cependant, dans ce dernier cas, la modélisation précise des transcrits demeure le point essentiel pour éviter les artefacts. Aussi, pour chaque espèce d'intérêt, les efforts en faveur d'une annotation du génome qui soit collective et partagée pour être la plus exhaustive et précise possible, telle que discutée dans les parties précédentes, est une brique de base essentielle à l'obtention de résultats biologique de qualité.

2. La symphonie génétique : l'orchestre caché des QTL et des eQTL

Afin d'étudier par GWAS les variants – en majorité régulateurs – impliqués dans les caractères complexes, un nombre important d'individus est nécessaire ainsi que de nombreux marqueurs bien répartis sur le génome et correctement génotypés pour la population d'intérêt. Dans le cadre de la sélection génétique ou pour le diagnostic médical, l'utilisation de puces de génotypage à basse ou moyenne densité est bien souvent la norme, car elles permettent, après un contrôle qualité, d'obtenir de tels marqueurs à bas coûts (voir Résultats §2.3). Cependant, ces puces sont avec a priori, elles sont construites en amont, à un instant donné et selon le génome de référence disponible à ce moment. Les marqueurs sont choisis pour être bien répartis dans le génome, mais aussi pour être polymorphes dans la majorité des populations d'intérêt. Comme montré dans le paragraphe « Résultats §2.1 », lorsque l'étude scientifique combine à la fois démarche GWAS et eQTL pour identifier les variant/gènes dans les signaux GWAS, les échantillons RNAseq alors disponibles apparaissent comme une alternative pour densifier le nombre de marqueurs (~500K) pour la population d'intérêt, tout en ciblant les zones exprimées du génome et donc potentiellement impactantes. De plus, cette alternative permet de considérer une version plus actualisée du génome de référence et en conséquence des variants dans des loci non couverts par les puces. Cela est illustré par la détection de nouveaux variants dans les micro-chromosomes récemment ajoutés lors du dernier assemblage GRCg7b de la poule (Introduction *review-2*) [114].

À l'issue des analyses GWAS, il est d'usage de délimiter une zone QTL d'intérêt qu'il convient alors d'explorer afin de prioriser d'éventuels gènes candidats. Cependant, la délimitation du QTL reste encore un sujet à débat. Si la valeur de 1 Mb (500 kb de part et d'autre du variant *leader*) est souvent retrouvée dans la littérature, des analyses portant sur la persistance du LD pour la population d'intérêt peuvent permettre de préciser la zone. Notons par ailleurs, que cette valeur peut varier selon la taille du chromosome comme chez la poule pour laquelle la persistance du DL apparait plus faible pour les micro-chromosomes [625]. Pour finir, il est également possible d'observer des zones de rupture, où un ensemble de marqueurs n'est pas en LD alors que d'autres plus loin le sont, permettant ainsi d'évincer certaines régions chromosomiques (voir Résultats §3.2).

Une fois cette zone délimitée, il est courant de s'intéresser aux PCG présents dans le QTL, car ces gènes sont les mieux modélisés et leurs fonctions sont mieux caractérisées. Il est ainsi possible d'identifier des gènes candidats causaux de par leur fonction biologique en lien avec le caractère d'intérêt. Ainsi, la majorité des GWAS portant sur les espèces d'élevage se résume à des régions QTL identifiées et des gènes candidats PCG proposés pour chacune d'entre elles [417]. En trouver apparait alors relativement fréquent au vu du nombre de gènes se situant dans les 1 Mb autour du variant *leader*. Rappelons que chez la poule dont le génome avoisine les 1 Gb, 20 000 PCG sont identifiés, ce qui représente environ 20 PCG dans une fenêtre de 1 MB.

Dans certaines études, portant sur les espèces modèles ou chez l'humain pour lesquels les exomes sont disponibles, les variants des parties codantes sont souvent disponibles. Il est alors facile de prédire leurs impacts sur les séquences protéiques [517]. Cet impact est prédit grâce au code génétique, mais peut également être quantifié par des études de conservation des acides aminées entre différentes espèces (*e.g.*, SIFT score [518]). De plus, il est supposé que les effets de ces variants sont potentiellement les plus importants, car affectant la structure de la protéine (et non sa quantité) et sont donc plus simples à détecter. Notons cependant que la prédiction de l'impact est parfois faussée et qu'il convient de prendre en compte la présence de plusieurs variants (MNV) au sein d'un même codon et porté par le même chromosome (Résultats §2.2) [626]. Ces travaux ont par ailleurs montré que la majorité des MNV initialement prédits avec un effet fort (*stop gained* ou *missense variant* avec un SIFT score faible) présentaient une diminution de l'impact de la prédiction.
En résumé, l'analyse des PCG, et en particulier des parties codantes, est bien souvent la première stratégie appliquée, d'où la surestimation de leur représentation dans la littérature par rapport aux couples variants régulateurs/gènes régulés ou aux gènes régulateurs qui sont, jusqu'à présent, très peu identifiés, car plus difficile à détecter comme discuté ci-après.

Parmi les gènes d'une région GWAS d'intérêt, il convient donc de s'intéresser également aux gènes régulateurs et notamment aux lncRNA que l'on sait en grand nombre (Résultats §1.1 et 1.2) [124, 125] et qui peuvent également expliquer, en partie, la variation du caractère d'intérêt. Cependant, à la différence des PCG, on ne sait actuellement pas prédire les impacts

des variants sur la structure des lncRNA. En effet, les règles tirées du code génétique pour prédire l'impact d'un variant sur la fonction du transcrit associé au gène long non codant ne peuvent être appliquées. Par ailleurs, en subissant une pression de sélection moins forte que les PCG, le nombre de variants dans les exons des lncRNA apparait plus élevé que pour les PCG et leur séquence primaire se trouve alors peu conservée entre espèces [276, 283]. Il apparait ainsi complexe de prioriser les variants à l'échelle du nucléotide par conservation entre espèces. Une première approche serait donc de réaliser de la prédiction d'impact en détectant les variants présents dans des k-mers fonctionnels (Discussion §1.3). À l'instar du SIFT, et en supposant des lncRNA orthologues entre plusieurs espèces (Résultats §1.4), il serait alors possible d'observer si des variations sont tolérées ou non dans d'autres espèces, indicateur de l'effet du variant sur la fonction du lncRNA. Une autre voie de prédiction des variants serait également de quantifier l'impact d'un variant sur la configuration spatiale des lncRNA, supposée conditionner sa fonction. La prédiction des effets des variants pourrait alors être quantifié au travers de leur impact sur la stabilité de la structure thermodynamique du lncRNA, cependant cela nécessite une puissance de calcul importante pour une étude systématique [627, 628].

La limite des approches précédentes portant sur les variants présents dans les gènes et qu'elles ne s'intéressent qu'aux modifications de la structure des transcrits et des protéines, respectivement pour les gènes de type lncRNA et PCG, or une majorité des variations associées aux caractères est située dans des régions non codantes régulatrices et dont les effets sur la régulation des gènes sont généralement inconnus. L'une des approches mise en place a alors été de cartographier les locus responsables de la variabilité de caractères quantitatifs moléculaires (molQTL) tels que : *i)* l'expression des gènes (eQTL et lncQTL pour les gènes PCG et lncRNA respectivement), *ii)* l'expression des exons (exQTL) et notamment de l'exon 3'UTR qui est le reflet de mécanismes variés de dégradation de l'ARN par l'extrémité 3'UTR ou site de polyadénylation (3'aQTL), ou encore *iii)* l'épissage (sQTL).

Les études à grande échelle comme les GTEx (Résultats §3.1) ont ainsi produit (cas du GTEx humain) [389] ou agrégé (cas des GTEx bovin, porc et poule) [405, 406, 574] un riche ensemble de données, afin d'analyser ces molQTL. Ces analyses GTEx consistent à associer les génotypages sur un grand nombre de marqueurs avec différents phénotypes moléculaires

quantifiés dans quelques dizaines de tissus (20 à 50), sur des dizaines, voire des centaines d'individus non apparentés. D'un point de vue pratique, alors que les études pilotes ont tendance à comporter entre 50 et 200 individus par tissu, quelques centaines d'individus sont généralement nécessaires pour obtenir une bonne puissance statistique dans une étude molQTL standard. Enfin, le panel de tissus analysés dépasse en général les 20 tissus afin de couvrir un nombre important de processus biologiques. La collecte rigoureuse des métadonnées associées aux échantillons biologiques (individu *x* tissu) est un élément critique pour le contrôle de la qualité des données (bonne assignation) et l'identification des sources potentielles de variation des données lors de la sélection des covariables dans les modèles d'analyse GWAS.

L'ensemble de ces critères a été considéré dans l'étude pilote QTL-eQTL effectuée au laboratoire sur une population de poules pondeuses commerciales (Résultats §3.2) et qui va se poursuivre lors d'une nouvelle thèse. L'objectif étant de comprendre la composante génétique en partie responsable de caractères liés au métabolisme des acides gras à âge avancé (régulations de gènes PCG impliqués dans ce métabolisme, stéatose hépatique, gras corporel…) et au vu du coût d'obtention d'un transcriptome tissulaire par RNAseq, un seul tissu d'étude a été choisi. Ainsi les transcriptomes de foie d'environ 500 individus de 90 semaines ont été quantifiés par RNAseq et l'ensemble des métadonnées associés à chaque individu a été soigneusement renseignée.

Concernant la puissance statistique de détection des molQTL, le nombre de molQTL significatifs pour un nombre donné d'échantillons est relativement stable entre étude et augmente avec le nombre d'individus analysés. Dans le cas des eQTL, la majorité de ceux présentant un impact fort (*log2(aFC)* > 1) sont détectés à partir d'environ 200-300 échantillons. Notons qu'un panel de 100 individus permet de détecter approximativement 50 % des eQTL à impact fort [405]. Des dispositifs de plus grande taille permettent alors de découvrir des molQTL aux effets plus faibles, y compris des QTL secondaires, ainsi que des associations de variants à faible fréquence d'allèles plus robustes.

Quelles que soient les études GTEx, même si la cartographie des cis-molQTL est effectuée dans une large fenêtre, en général de +/- 1 Mb, les variants *leader* ont une localisation conforme à l'attendue, comme dans les promoteurs et *enhancer* pour les eQTL ou dans les sites de *splicing*

pour les sQTL. Cet enrichissement montre que suffisamment de recombinants à petite distance sont présents et que les effets sont assez forts pour que le variant observé comme *leader* soit assez proche du variant causal. De tels résultats étaient difficilement envisageables, les régions en LD étant sur des intervalles de l'ordre de la centaine de kilobases dans les populations animales et englobant donc en général région enhancer/promoteur, exons et sites de splicing. Cependant, la force des dispositifs GTEx est d'avoir agrégé de nombreuses populations permettant d'observer des régions en fort LD (r2 > 0,8) mais sur de petites zones, de l'ordre de quelques kilobases, permettant ainsi de distinguer différents types de régions régulatrices au sein d'un même gène. Il faut toutefois souligner que dans les études ne comportant qu'une population, le LD est tel que tous les variants en fort LD avec le variant leader (LD > 0,8) sont considérés comme proche du variant causal [629]. Ceci montre alors que ces approches appliquées à une seule population, comme dans notre cas, apportent rarement de l'information nouvelle quant aux régions régulatrices, mais permettent plutôt d'identifier les phénotypes moléculaires régulés, pouvant ensuite aider à prioriser les gènes causaux candidats dans les QTL GWAS.

Les études sur le partage des signaux pour différents molQTL donnent des résultats également conformes à la connaissance des grands types de régulation que l'on a de l'expression génique. Les exons étant des parties de transcrits de type PCG, ces derniers codant des protéines, il est cohérent de trouver une colocalisation importante (même si incomplète) entre les exQTL, eQTL et pQTL (les pQTL étant des QTL associés à variations de quantités des protéines) [630]. Les résultats sont différents lorsqu'il s'agit d'observer les sQTL et eQTL qui sont pour la plupart indépendants, de même pour les 3a'QTL et eQTL [405]. Il semblerait ainsi que des variants puissent réguler spécifiquement certains transcrits alternatifs (sQTL, 3'aQTL) sans affecter l'expression globale du gène (eQTL). Cette observation est due à la nature des phénotypes analysés. En effet, comme montré par Brotman et al. [629], les sQTL qui ne colocalisent pas avec les eQTL sont bien souvent associés à des transcrits alternatifs dont l'expression est significativement plus faible que l'expression globale du gène qui elle englobe l'expression de tous les transcrits. L'effet du sQTL est alors insuffisant pour expliquer la variation d'expression globale du gène et un autre variant souvent dans les promoteurs et plus impactant est donc associé au eQTL.

En résumé, ce jeu des colocalisation partielles ou absentes reflète la variété des mécanismes de la régulation de l'expression génique (prise au sens large) aux niveaux transcriptionnel (eQTL), post-transcriptionnel comme l'épissage (sQTL) et la dégradation des transcrits (3'aQTL), traductionnel et post-traductionnelle (pQTL).
Un dernier point à souligner concernant ces analyses GTEx est que le paysage régulateur peut dépendre du type cellulaire, un molQTL pouvant être actif dans un seul ou certains tissus ou dans l'ensemble des tissus où le gène est exprimé. Les différents travaux GTEx, quelles que soit les espèces, montrent que les tissus biologiquement proches ont tendance à se regrouper sur la base des effets des molQTL, suggérant donc des mécanismes de régulations partagés entre tissus biologiquement proches. Notons que ces études montrent également que les cis-sQTL et cis-lncQTL semblent être plus spécifiques des tissus que les cis-eQTL. Pour aller plus loin, un atlas sur données single-cell pourrait être intéressant à générer pour chaque tissu de façon à explorer de manière plus fine la régulation génique spécifique de populations cellulaires particulières du tissu d'intérêt. De plus, la quantification des proportions des populations cellulaires est d'ores et déjà rendu possible sur un grand nombre d'échantillons RNAseq via des approches de déconvolution *in silico* à partir d'un petit nombre de données *single cell* [591].

En conclusion, les résultats générés par les consortiums GTEX contribuent à améliorer la connaissance des éléments régulateurs du génome et viennent enrichir ceux déjà identifiés par d'autres projets, tel que le projet GENE-SWitCH [631] chez la poule qui utilise des méthodes de biologie moléculaire de type ChIPseq ou ATACseq pour annoter ces éléments. Les résultats du GTEx permettent également de mettre en relation des régions régulatrices (marquées par le variant *leader*) et le gène régulé, ce qui peut être utile, par exemple, pour associer aux gènes régulés les régions *enhancer* précédemment détectées.

Revenons maintenant aux signaux GWAS détectés pour l'étude des caractères complexes d'intérêt en faisant le lien avec ces molQTL régulant l'expression des gènes. Dans les différents papiers GTEx humain, bovin, porc et poule, une dernière partie est consacrée à la mise en lien entre les molQTL détectés et les signaux GWAS accumulés dans la littérature pour l'espèce d'intérêt. De manière globale, plus de 50% des signaux GWAS ne sont pas expliqués par les molQTL actuellement connus. Par exemple, pour 5 385 loci associés à 87 traits analysés dans

le cadre du projet GTEx humain, seuls 43 % des signaux GWAS étaient colocalisés avec des eQTL. Pour le ChickenGTEX, sur l'ensemble des 1 176 loci significatifs, 1 059 (90 %) ont pu être expliqués par au moins un molQTL dans un des 28 tissus. Cependant, soulignons qu'une colocalisation était considérée comme présente si elle était détectée significative dans au moins une des quatre méthodes appliquées.

Différentes hypothèses ont été formulées concernant ce nombre élevé de signaux GWAS ne colocalisant pas avec un molQTL dont une récemment, énoncé par Mostafavi et al., 2023 [632]. *i)* Les auteurs montrent que les gènes les plus proches des signaux GWAS (supposés ici être les gènes causaux) sont enrichis en gènes très contraints (donc a priori avec un rôle physiologique majeur), ayant de diverses fonctions (soulignées par un nombre élevé de termes GO), et/ou encore avec des actions de régulations variées (comme les facteurs de transcription) alors que ce n'est pas le cas pour les gènes régulés par les variants eQTL. Ainsi, à l'échelle moléculaire, la variation d'expression n'est pas suffisante pour détecter ces variants régulateurs (pas de signaux eQTL), tandis qu'à l'échelle de l'organisme, les actions cumulées de ces variants régulateurs, de par les multiples fonctions du gène régulé, conduisent à une variation suffisamment importante sur le caractère pour être détecté (signaux GWAS). *ii)* Une autre hypothèse pouvant expliquer le nombre limité de colocalisation entre GWAS et eQTL est que les dispositifs d'individus utilisés par le GTEx et les GWAS sont disjoints. Ainsi, un eQTL peut ne pas être détecté si la variation d'expression n'est observée que dans un contexte biologique spécifique, non présent dans le dispositif GTEx (*e.g.*, un stade physiologique), ou encore si le variant régulateur associé au signal GWAS est absent des populations utilisées dans le GTEx. Aussi, il est intéressant de mettre en œuvre dans un même dispositif les approches GWAS et eQTL comme nous le faisons avec le dispositif « poules pondeuses » (voir Résultats §2.3) permettant ainsi d'observer les signaux GWAS et eQTL dans les mêmes conditions génétiques (même population) et biologique (même stade physiologique, d'alimentation, même sexe). L'intérêt d'un tel dispositif a déjà été illustré via les travaux de l'équipe sur le gène BMCO1 et la couleur de la viande [633] où le signal GWAS détecté dans une population de poulet de chair co-localisait avec un eQTL dans le muscle, régulant un gène codant une enzyme clé de la conversion du b-carotène en rétinal incolore. Des études expérimentales complémentaires ont alors permis de renforcer le statut causal de ce gène BCMO1 dans le caractère d'intérêt, en particulier en identifiant 2 SNP dans le promoteur

responsable de la variation d'expression de ce gène. Notons qu'une faiblesse d'un tel dispositif est la taille des régions en LD qui est plus grande que dans le GTEx, ne permettant donc pas de localiser avec précision le variant régulateur.

En conclusion, si les approches basées sur l'étude des variants et de leurs impacts sur l'expression des gènes avaient pour volonté première d'aider à la compréhension et au décryptage des signaux GWAS associés aux caractères complexes en se positionnant notamment à l'interface entre le génotype et le phénotype, elles permettent surtout à ce jour d'enrichir la connaissance des séquences régulatrices et leurs liens avec les gènes régulés. Ces travaux soulignent par ailleurs la diversité et la modularité (action dans ces conditions spécifiques) de ces régulateurs qui agissent à différentes étapes de la régulation de l'expression d'un gène. Ainsi, s'il est tout de même possible de prioriser de potentiels gènes candidats causaux (voire variants), il est fondamental de garder son intégrité scientifique en ne tirant pas de conclusions hâtives, le retour aux expérimentations de biologie moléculaire étant inévitable pour valider le statut de gène/variant causal.

Bibliographie

1. CNRTL. Hérédité : étymologie de hérédité. 2023. https://www.cnrtl.fr/etymologie/h%C3%A9r%C3%A9dit%C3%A9. Accessed 28 Jul 2023.

2. Zirkle C. The Inheritance of Acquired Characters and the Provisional Hypothesis of Pangenesis. The American Naturalist. 1935;69:417–45.

3. Hippocrates. Hippocratic writings. Harmondsworth ; New York : Penguin; 1978.

4. Kalachanis K, Michailidis I. The Hippocratic View on Humors and Human Temperament. 2015;2:1–5.

5. Jouanna J, Allies N. The Legacy of the Hippocratic Treatise the Nature of Man: The Theory of the Four Humours. In: van der Eijk P, editor. Greek Medicine from Hippocrates to Galen. Brill; 2012. p. 335–60.

6. Cobb M. Heredity before genetics: a history. Nat Rev Genet. 2006;7:953–8.

7. Noble D. Chapter 21 - Exosomes, gemmules, pangenesis and Darwin. In: Edelstein L, Smythies J, Quesenberry P, Noble D, editors. Exosomes. Academic Press; 2020. p. 487–501.

8. Defradas J. 140. Aristote. De la génération des animaux. Texte établi et traduit par Pierre Louis (Collection des Universités de France). Paris, Les Belles-Lettres, 1961. Revue des Études Grecques. 1963;76:481–2.

9. Bonnard J-B. « Il paraît en effet que les fils ressemblent aux pères ». In: Prost F, Wilgaux J, editors. Penser et représenter le corps dans l'Antiquité. Rennes: Presses universitaires de Rennes; 2015. p. 307–18.

10. Brumbaugh RS. Plato's genetic theory. Journal of Heredity. 1954;45:191–6.

11. Thiery M. Reinier De Graaf (1641–1673) and the Graafian follicle. Gynecol Surg. 2009;6:189–91.

12. Lane N. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals.' Philosophical Transactions of the Royal Society B: Biological Sciences. 2015;370:20140344.

13. Henig RM. The monk in the garden : the lost and found genius of Gregor Mendel, the father of genetics. Boston : Houghton Mifflin; 2000.

14. Vecerek O. Johann Gregor Mendel as a Beekeeper. Bee World. 1965;46:86–96.

15. Mendel G. Experiments in plant hybridization. 1865.

16. Darnell JE, Lodish HF, Baltimore D. Molecular Cell Biology. Spektrum Akademischer Verlag; 1986.

17. Reid JB, Ross JJ. Mendel's Genes: Toward a Full Molecular Characterization. Genetics. 2011;189:3–10.

18. Magner LN. A History of the Life Sciences, Revised and Expanded. CRC Press; 2002.

19. Hartl DL, Orel V. What Did Gregor Mendel Think He Discovered? Genetics. 1992;131:245–53.

20. Robert Bear, David Rintoul, Bruce Snyder, Martha Smith-Caldas, Christopher Herren, Eva Horne. Laws of inheritance. In: Principles of Biology. OpenStax CNX; 2023.

21. Porter TM. The curious case of blending inheritance. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences. 2014;46:125–32.

22. Edwards A. Punnett's square. Studies in history and philosophy of biological and biomedical sciences. 2012;43:219–24.

23. Galton DJ. Did Mendel falsify his data? QJM: An International Journal of Medicine. 2012;105:215–6.

24. Kottler MJ. Hugo de Vries and the rediscovery of Mendel's laws. Annals of Science. 1979;36:517–38.

25. Rheinberger H-J. Re-discovering Mendel: The Case of Carl Correns. Sci & Educ. 2015;24:51–60.

26. Monaghan FV, Corcos AF. Tschermak: a non-discoverer of mendelism. II. A critique. J Hered. 1987;78:208–10.

27. Lenay C. Hugo De Vries: from the theory of intracellular pangenesis to the rediscovery of Mendel. Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie. 2000;323:1053–60.

28. Dahm R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum Genet. 2008;122:565–81.

29. Deichmann U. Chromatin: Its history, current research, and the seminal researchers and their philosophy. Perspectives in Biology and Medicine. 2015;58:143–64.

30. Flemming W. Zellsubstanz, Kern und Zelltheilung. Vogel; 1882.

31. Cremer T, Cremer C. Centennial of Wilhelm Waldeyer's introduction of the term "chromosome" in 1888. Cytogenetics and Cell Genetics. 2008;48:66–7.

32. Zacharias H. Key word: Chromosome. Chromosome Res. 2001;9:345–55.

33. Paweletz N. Walther Flemming: pioneer of mitosis research. Nat Rev Mol Cell Biol. 2001;2:72–5.

34. Clare O'Connor, Ilona Miko. Developing the Chromosome Theory | Learn Science at Scitable. Nature Education. 2008.

35. Crow EW, Crow JF. 100 Years Ago: Walter Sutton and the Chromosome Theory of Heredity. Genetics. 2002;160:1–4.

36. Laubichler MD, Davidson EH. Boveri's long experiment: Sea urchin merogones and the establishment of the role of nuclear chromosomes in development. Dev Biol. 2008;314:1–11.

37. Cremer T, Cremer M. Chromosome Territories. Cold Spring Harb Perspect Biol. 2010;2:a003889.

38. Sutton WS. On the morphology of the chromosom group in Brachystola magna. The Biological Bulletin. 1902;4:24–39.

39. Carey SB, Aközbek L, Harkess A. The contributions of Nettie Stevens to the field of sex chromosome biology. Philosophical Transactions of the Royal Society B: Biological Sciences. 2022;377:20210215.

40. Wikipédia - Travail collaboratif. Nettie Stevens. Wikipédia. 2023.

41. Carothers EE. The Maturation Divisions in Relation to the Segregation of Homologous Chromosomes. The Quarterly Review of Biology. 1926;1:419–35.

42. Johannsen W. Elemente der exakten erblichkeitslehre. Deutsche wesentlich erweiterte ausgabe in fünfundzwanzig vorlesungen. Jena: G. Fischer; 1909.

43. Ruelland JG. 6. Le gène et les déterminismes. In: L'empire des gènes : Histoire de la sociobiologie. Lyon: ENS Éditions; 2014. p. 95–116.

44. Morgan TH. Sex Limited Inheritance in Drosophila. Science. 1910;32:120–2.

45. Morgan TH. Complete Linkage in the Second Chromosome of the Male of Drosophila. Science. 1912;36:719–20.

46. Sturtevant AH. Genetic studies on drosophila simulans. Hybrids with drosophila melanogaster. Genetics. 1920;5:488–500.

47. Sturtevant AH. A Case of Rearrangement of Genes in Drosophila1. Proceedings of the National Academy of Sciences. 1921;7:235–7.

48. Sumner JB. The isolation and crystallization of the enzyme urease. Journal of Biological Chemistry. 1926;69:435–41.

49. Northrop JH. Crystalline pepsin. J Gen Physiol. 1930;13:739–66.

50. Shampo MA, Kyle RA. John Northrop—Definitive Study of Enzymes. Mayo Clinic Proceedings. 2000;75:254.

51. Griffith F. The Significance of Pneumococcal Types. J Hyg (Lond). 1928;27:113–59.

52. Avery OT, Macleod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med. 1944;79:137–58.

53. Amsterdamska O. From pneumonia to DNA: the research career of Oswald T. Avery. Hist Stud Phys Biol Sci. 1993;24 pt 1:1–40.

54. Hershey AD, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol. 1952;36:39–56.

55. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171:737–8.

56. Klug A. Rosalind Franklin and the Discovery of the Structure of DNA. Nature. 1968;219:808–10.

57. Crick FH. On protein synthesis. Symp Soc Exp Biol. 1958;12:138–63.

58. Nirenberg M. Historical review: Deciphering the genetic code--a personal account. Trends Biochem Sci. 2004;29:46–54.

59. Nirenberg MW, Matthaei JH. The dependence of cell- free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A. 1961;47:1588–602.

60. Khorana HG, Büuchi H, Ghosh H, Gupta N, Jacob TM, Kössel H, et al. Polynucleotide Synthesis and the Genetic Code. Cold Spring Harb Symp Quant Biol. 1966;31:39–49.

61. Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. Cell. 1977;12:1–8.

62. Berget SM, Moore C, Sharp PA. Spliced segments at the 5’ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A. 1977;74:3171–5.

63. Baltimore D. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. Nature. 1970;226:1209–11.

64. Temin HM, Mizutani S. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. Nature. 1970;226:1211–3.

65. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961;3:318–56.

66. Jacob F, Perrin D, Sanchez C, Monod J. [Operon: a group of genes with the expression coordinated by an operator]. C R Hebd Seances Acad Sci. 1960;250:1727–9.

67. Kresge N, Simoni RD, Hill RL. The Discovery of tRNA by Paul C. Zamecnik. Journal of Biological Chemistry. 2005;280:e37–9.

68. Hoagland MB, Zamecnik PC, Stephenson ML. Intermediate reactions in protein biosynthesis. Biochimica et Biophysica Acta. 1957;24:215–6.

69. Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011;470:187–97.

70. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

71. Williams N. Yeast genome sequence ferments new research. Science. 1996;272:481.

72. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of Drosophila melanogaster. Science. 2000;287:2185–95.

73. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282:2012–8.

74. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–815.

75. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432:695–716.

76. Ritchie H, Rosado P, Roser M. Meat and Dairy Production. Our World in Data. 2017.

77. Ritchie H, Rosado P, Roser M. Environmental Impacts of Food Production. Our World in Data. 2022.

78. Mozdziak PE, Petitte JN. Status of transgenic chicken models for developmental biology. Dev Dyn. 2004;229:414–21.

79. Kumar S, Hedges SB. A molecular timescale for vertebrate evolution. Nature. 1998;392:917–20.

80. Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M. Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes. 2019;12:315.

81. Omenn GS. Reflections on the HUPO Human Proteome Project, the Flagship Project of the Human Proteome Organization, at 10 Years. Mol Cell Proteomics. 2021;20:100062.

82. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33 Suppl:245–54.

83. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270:467–70.

84. Weber APM. Discovering New Biology through Sequencing of RNA1. Plant Physiol. 2015;169:1524–31.

85. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science. 2008;320:1344–9.

86. Biémont C, Vieira C. Genetics: junk DNA as an evolutionary force. Nature. 2006;443:521–4.

87. Kidwell MG, Lisch DR. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution. 2001;55:1–24.

88. Medstrand P, van de Lagemaat LN, Dunn CA, Landry J-R, Svenback D, Mager DL. Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res. 2005;110:342–52.

89. Hesper B, Hogeweg P. Bio-informatics: a working concept. A translation of “Bio-informatica: een werkconcept” by B. Hesper and P. Hogeweg. 2021.

90. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology. PLoS Computational Biology. 2011;7.

91. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.

92. Kneale GG, Kennard O. The EMBL nucleotide sequence data library. Biochem Soc Trans. 1984;12:1011–4.

93. Burks C, Fickett JW, Goad WB, Kanehisa M, Lewitter FI, Rindone WP, et al. The GenBank nucleic acid sequence database. Comput Appl Biosci. 1985;1:225–33.

94. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

95. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227:1435–41.

96. Burge CB, Karlin S. Finding the genes in genomic DNA. Curr Opin Struct Biol. 1998;8:346–54.

97. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268:78–94.

98. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25:1915–27.

99. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012;22:1775–89.

100. Kung JTY, Colognori D, Lee JT. Long Noncoding RNAs: Past, Present, and Future. Genetics. 2013;193:651–69.

101. Ponting CP, Haerty W. Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. Annu Rev Genomics Hum Genet. 2022;23:153–72.

102. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen L-L, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023;24:430–47.

103. Lagarrigue S, Lorthiois M, Degalez F, Gilot D, Derrien T. LncRNAs in domesticated animals: from dog to livestock species. Mamm Genome. 2022;33:248–70.

104. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The Transcriptional Landscape of the Mammalian Genome. Science. 2005;309:1559–63.

105. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.

106. Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. Nat Struct Mol Biol. 2015;22:5–7.

107. Snyder MP, Gingeras TR, Moore JE, Weng Z, Gerstein MB, Ren B, et al. Perspectives on ENCODE. Nature. 2020;583:693–8.

108. Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, et al. Long noncoding RNA repertoire in chicken liver and adipose tissue. Genet Sel Evol. 2017;49:6.

109. Hüttenhofer A, Schattner P, Polacek N. Non-coding RNAs: hope or hype? Trends in Genetics. 2005;21:289–97.

110. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:10.1146/annurev-biochem-051410–92902.

111. Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 2007;17:556–65.

112. Zhao L, Wang J, Li Y, Song T, Wu Y, Fang S, et al. NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Res. 2021;49:D165–71.

113. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. Nat Rev Genet. 2018;19:535–48.

114. Smith J, Alfieri JM, Anthony N, Arensburger P, Athrey GN, Balacco J, et al. Fourth Report on Chicken Genes and Chromosomes 2022. Cytogenetic and Genome Research. 2023;162:405–528.

115. Bridges MC, Daulagala AC, Kourtidis A. LNCcation: lncRNA localization and function. Journal of Cell Biology. 2021;220:e202009045.

116. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.

117. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.

118. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45:W12–6.

119. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011;27:i275–82.

120. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004;32 Database issue:D115–9.

121. Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics. 2014;15:311.

122. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. 2017;45:e57.

123. Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, et al. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. Genome Biology. 2016;17:14.

124. Jehl F, Muret K, Bernard M, Boutin M, Lagoutte L, Désert C, et al. An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. Sci Rep. 2020;10:20457.

125. Degalez F, Charles M, Foissac S, Zhou H, Guan D, Fang L, et al. Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues. 2023;:2023.08.18.553750.

126. Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature. 2013;493:231–5.

127. Sydney School of Veterinary Science, University of Sydney. OMIA - Online Mendelian Inheritance in Animals. 2023. https://www.omia.org/home/. Accessed 21 Sep 2023.

128. Dong J, He C, Wang Z, Li Y, Li S, Tao L, et al. A novel deletion in KRT75L4 mediates the frizzle trait in a Chinese indigenous chicken. Genetics Selection Evolution. 2018;50:68.

129. Chen B, Xi S, El-Senousey HK, Zhou M, Cheng D, Chen K, et al. Deletion in KRT75L4 linked to frizzle feather in Xiushui Yellow Chickens. Anim Genet. 2022;53:101–7.

130. Swiezewski S, Liu F, Magusin A, Dean C. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature. 2009;462:799–802.

131. Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, et al. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science. 2010;329:689–93.

132. Desert C, Baéza E, Aite M, Boutin M, Le Cam A, Montfort J, et al. Multi-tissue transcriptomic study reveals the main role of liver in the chicken adaptive response to a switch in dietary energy source through the transcriptional regulation of lipogenesis. BMC Genomics. 2018;19:187.

133. Derrien T. FEELnc : 2023.

134. Ponting CP, Oliver PL, Reik W. Evolution and Functions of Long Noncoding RNAs. Cell. 2009;136:629–41.

135. Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y, Lawrence J, et al. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell. 1992;71:527–42.

136. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, et al. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell. 1992;71:515–26.

137. Yap KL, Li S, Muñoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, et al. Molecular Interplay of the Non-coding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. Mol Cell. 2010;38:662–74.

138. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature. 2012;491:454–7.

139. Shibayama Y, Fanucchi S, Magagula L, Mhlanga MM. lncRNA and gene looping. Transcription. 2014;5:e28658.

140. Wang C, Jia L, Wang Y, Du Z, Zhou L, Wen X, et al. Genome-wide interaction target profiling reveals a novel Peblr20-eRNA activation pathway to control stem cell pluripotency. Theranostics. 2020;10:353–70.

141. Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. Long non-coding RNAs with enhancer-like function in human. Cell. 2010;143:46–58.

142. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, et al. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature. 2013;494:497–501.

143. Setten RL, Chomchan P, Epps EW, Burnett JC, Rossi JJ. CRED9: a differentially expressed elncRNA regulates expression of transcription factor CEBPA. RNA. 2021;27:891–906.

144. Chowdhury IH, Narra HP, Sahni A, Khanipov K, Fofanov Y, Sahni SK. Enhancer Associated Long Non-coding RNA Transcription and Gene Regulation in Experimental Models of Rickettsial Infection. Frontiers in Immunology. 2019;9.

145. Schmitz K-M, Mayer C, Postepska A, Grummt I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes Dev. 2010;24:2264–9.

146. Dong L, Ni J, Hu W, Yu C, Li H. Upregulation of Long Non-Coding RNA PlncRNA-1 Promotes Metastasis and Induces Epithelial-Mesenchymal Transition in Hepatocellular Carcinoma. Cellular Physiology and Biochemistry. 2016;38:836–46.

147. Uesaka M, Agata K, Oishi T, Nakashima K, Imamura T. Evolutionary acquisition of promoter-associated non-coding RNA (pancRNA) repertoires diversifies species-dependent gene activation mechanisms in mammals. BMC Genomics. 2017;18:285.

148. Flynn RA, Almada AE, Zamudio JR, Sharp PA. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proceedings of the National Academy of Sciences. 2011;108:10460–5.

149. Nair L, Chung H, Basu U. Regulation of long non-coding RNAs and genome dynamics by the RNA surveillance machinery. Nat Rev Mol Cell Biol. 2020;21:123–36.

150. Wang Y, Yao J, Meng H, Yu Z, Wang Z, Yuan X, et al. A novel long non-coding RNA, hypoxia-inducible factor-2α promoter upstream transcript, functions as an inhibitor of osteosarcoma stem cells in vitro. Molecular Medicine Reports. 2015;11:2534–40.

151. Feng J, Bi C, Clark BS, Mady R, Shah P, Kohtz JD. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev. 2006;20:1470–84.

152. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved Elements in the Human Genome. Science. 2004;304:1321–5.

153. Gibert MK, Sarkar A, Chagari B, Roig-Laboy C, Saha S, Bednarek S, et al. Transcribed Ultraconserved Regions in Cancer. Cells. 2022;11:1684.

154. Braconi C, Valeri N, Kogure T, Gasparini P, Huang N, Nuovo GJ, et al. Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. Proceedings of the National Academy of Sciences. 2011;108:786–91.

155. Chillón I, Marcia M. The molecular structure of long non-coding RNAs: emerging patterns and functional implications. Critical Reviews in Biochemistry and Molecular Biology. 2020;55:662–90.

156. GENCODE. GENCODE - Biotypes. https://www.gencodegenes.org/pages/biotypes.html. Accessed 24 Aug 2023.

157. EMBL-EBI. International Nucleotide Sequence Database Collaboration. https://www.insdc.org/submitting-standards/ncrna-vocabulary/. Accessed 24 Aug 2023.

158. Murphy M, Brown G, Wallin C, Tatusova T, Pruitt K, Murphy T, et al. Gene Help: Integrated Access to Genes of Genomes in the Reference Sequence Collection. In: Gene Help [Internet]. National Center for Biotechnology Information (US); 2022.

159. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet. 2016;17:47–62.

160. Muret K, Désert C, Lagoutte L, Boutin M, Gondret F, Zerjal T, et al. Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. BMC Genomics. 2019;20:882.

161. Wang X, Liu C, Zhang S, Yan H, Zhang L, Jiang A, et al. N6-methyladenosine modification of MALAT1 promotes metastasis via reshaping nuclear speckles. Dev Cell. 2021;56:702-715.e8.

162. Yamazaki T, Souquere S, Chujo T, Kobelke S, Chong YS, Fox AH, et al. Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. Mol Cell. 2018;70:1038-1053.e7.

163. Barral A, Déjardin J. Telomeric Chromatin and TERRA. Journal of Molecular Biology. 2020;432:4244–56.

164. Grote P, Wittler L, Hendrix D, Koch F, Währisch S, Beisaw A, et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev Cell. 2013;24:206–14.

165. Gibbons HR, Shaginurova G, Kim LC, Chapman N, Spurlock CF, Aune TM. Divergent lncRNA GATA3-AS1 Regulates GATA3 Transcription in T-Helper 2 Cells. Front Immunol. 2018;9:2512.

166. Kajino T, Shimamura T, Gong S, Yanagisawa K, Ida L, Nakatochi M, et al. Divergent lncRNA MYMLR regulates MYC by eliciting DNA looping and promoter-enhancer interaction. The EMBO Journal. 2019;38:e98441.

167. Dijk M van, Visser A, Buabeng KML, Poutsma A, Schors RC van der, Oudejans CBM. Mutations within the LINC-HELLP non-coding RNA differentially bind ribosomal and RNA splicing complexes and negatively affect trophoblast differentiation. HUM MOL GENET. 2015;24:5475–85.

168. Yu Y, Nangia-Makker P, Farhana L, Majumdar APN. A novel mechanism of lncRNA and miRNA interaction: CCAT2 regulates miR-145 expression by suppressing its maturation process in colon cancer cells. Molecular Cancer. 2017;16:155.

169. Guo J, Fang W, Sun L, Lu Y, Dou L, Huang X, et al. Ultraconserved element uc.372 drives hepatic lipid accumulation by suppressing miR-195/miR4668 maturation. Nat Commun. 2018;9:612.

170. Wang Y, Sun L, Wang L, Liu Z, Li Q, Yao B, et al. Long non-coding RNA DSCR8 acts as a molecular sponge for miR-485-5p to activate Wnt/β-catenin signal pathway in hepatocellular carcinoma. Cell Death Dis. 2018;9:1–13.

171. Zhu L, Wei Q, Qi Y, Ruan X, Wu F, Li L, et al. PTB-AS, a Novel Natural Antisense Transcript, Promotes Glioma Progression by Improving PTBP1 mRNA Stability with SND1. Mol Ther. 2019;27:1621–37.

172. Xu T -p, Liu X -x, Xia R, Yin L, Kong R, Chen W -m, et al. SP1-induced upregulation of the long noncoding RNA TINCR regulates cell proliferation and apoptosis by affecting KLF2 mRNA stability in gastric cancer. Oncogene. 2015;34:5648–61.

173. Sun Q, Song YJ, Prasanth KV. One locus with two roles: microRNA-independent functions of microRNA-host-gene locus-encoded long noncoding RNAs. Wiley Interdiscip Rev RNA. 2021;12:e1625.

174. Agranat-Tamir L, Shomron N, Sperling J, Sperling R. Interplay between pre-mRNA splicing and microRNA biogenesis within the supraspliceosome. Nucleic Acids Res. 2014;42:4640–51.

175. Morenos L, Chatterton Z, Ng JL, Halemba MS, Parkinson-Bates M, Mechinaud F, et al. Hypermethylation and down-regulation of DLEU2 in paediatric acute myeloid leukaemia independent of embedded tumour suppressor miR-15a/16-1. Molecular Cancer. 2014;13:123.

176. Wang H, Iacoangeli A, Popp S, Muslimov IA, Imataka H, Sonenberg N, et al. Dendritic BC1 RNA: Functional Role in Regulation of Translation Initiation. J Neurosci. 2002;22:10232–41.

177. Yu X, Zheng Q, Zhang Q, Zhang S, He Y, Guo W. MCM3AP-AS1: An Indispensable Cancer-Related LncRNA. Front Cell Dev Biol. 2021;9:752718.

178. Munschauer M, Nguyen CT, Sirokman K, Hartigan CR, Hogstrom L, Engreitz JM, et al. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. Nature. 2018;561:132–6.

179. Deng X, Li S, Kong F, Ruan H, Xu X, Zhang X, et al. Long noncoding RNA PiHL regulates p53 protein stability through GRWD1/RPL11/MDM2 axis in colorectal cancer. Theranostics. 2020;10:265–80.

180. Yan C, Chen J, Chen N. Long noncoding RNA MALAT1 promotes hepatic steatosis and insulin resistance by increasing nuclear SREBP-1c protein stability. Sci Rep. 2016;6:22640.

181. Polycarpou-Schwarz M, Groß M, Mestdagh P, Schott J, Grund SE, Hildenbrand C, et al. The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. Oncogene. 2018;37:4750–68.

182. Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science. 2016;351:271–5.

183. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science. 2005;309:1570–3.

184. Wang L, Li J, Zhou H, Zhang W, Gao J, Zheng P. A novel lncRNA Discn fine-tunes replication protein A (RPA) availability to promote genomic stability. Nat Commun. 2021;12:5572.

185. Pan L, Liang W, Fu M, Huang Z-H, Li X, Zhang W, et al. Exosomes-mediated transfer of long noncoding RNA ZFAS1 promotes gastric cancer progression. J Cancer Res Clin Oncol. 2017;143:991–1004.

186. Chen L, Yang W, Guo Y, Chen W, Zheng P, Zeng J, et al. Exosomal lncRNA GAS5 regulates the apoptosis of macrophages and vascular endothelial cells in atherosclerosis. PLOS ONE. 2017;12:e0185406.

187. Gil N, Ulitsky I. Regulation of gene expression by cis-acting long non-coding RNAs. Nat Rev Genet. 2020;21:102–17.

188. Statello L, Guo C-J, Chen L-L, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021;22:96–118.

189. Andergassen D, Rinn JL. From genotype to phenotype: genetics of mammalian long non-coding RNAs in vivo. Nat Rev Genet. 2022;23:229–43.

190. Schmitt AM, Chang HY. Long Noncoding RNAs in Cancer Pathways. Cancer Cell. 2016;29:452–63.

191. Li K, Wang Z. lncRNA NEAT1: Key player in neurodegenerative diseases. Ageing Research Reviews. 2023;86:101878.

192. Kohlmaier A, Holdt LM, Teupser D. Long noncoding RNAs in cardiovascular disease. Curr Opin Cardiol. 2023;38:179–92.

193. Karimi B, Dehghani Firoozabadi A, Peymani M, Ghaedi K. Circulating long noncoding RNAs as novel bio-tools: Focus on autoimmune diseases. Human Immunology. 2022;83:618–27.

194. Hill WG. Is Continued Genetic Improvement of Livestock Sustainable? Genetics. 2016;202:877–81.

195. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biology. 2015;16:57.

196. Giuffra E, Tuggle CK, FAANG Consortium. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annu Rev Anim Biosci. 2019;7:65–88.

197. Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Pérez-Enciso M. Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. BMC Genomics. 2011;12:552.

198. Li T, Wang S, Wu R, Zhou X, Zhu D, Zhang Y. Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. Genomics. 2012;99:292–8.

199. NIH. International Protein Nomenclature Guidelines. 2020. https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/. Accessed 24 Aug 2023.

200. Bruford EA, Braschi B, Denny P, Jones TEM, Seal RL, Tweedie S. Guidelines for Human Gene Nomenclature. Nat Genet. 2020;52:754–8.

201. Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, et al. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature. 1991;349:38–44.

202. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs. Cell. 2007;129:1311–23.

203. Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, et al. Genenames.org: the HGNC resources in 2023. Nucleic Acids Research. 2023;51:D1003–9.

204. International Committee on Standardized Genetic Nomenclature for Mice. MGI-Guidelines for Nomenclature of Genes, Genetic Markers, Alleles, & Mutations in Mouse & Rat. 2021. https://www.informatics.jax.org/mgihome/nomen/gene.shtml. Accessed 24 Aug 2023.

205. Rat Genome Nomenclature Committee. RGD: Help: Nomenclature Resources. https://rgd.mcw.edu/nomen/nomen.shtml. Accessed 24 Aug 2023.

206. Jones TEM, Yates B, Braschi B, Gray K, Tweedie S, Seal RL, et al. The VGNC: expanding standardized vertebrate gene nomenclature. Genome Biology. 2023;24:115.

207. Burt DW, Carrë W, Fell M, Law AS, Antin PB, Maglott DR, et al. The chicken gene nomenclature committee report. BMC Genomics. 2009;10:S5.

208. Zebrafish Nomenclature Committee. ZFIN Zebrafish Nomenclature Conventions - General Information - Confluence. 2022. https://zfin.atlassian.net/wiki/spaces/general/pages/1818394635/ZFIN+Zebrafish+Nomenclature+Conventions#ZFINZebrafishNomenclatureGuidelines-1. Accessed 24 Aug 2023.

209. Bruford EA. Highlights of the “Gene Nomenclature Across Species” Meeting. Human Genomics. 2010;4:213.

210. Wright MW, Bruford EA. Naming “junk”: Human non-protein coding RNA (ncRNA) gene nomenclature. Hum Genomics. 2011;5:90–8.

211. Wright MW. A short guide to long non-coding RNA gene nomenclature. Hum Genomics. 2014;8:7.

212. Seal RL, Chen L, Griffiths-Jones S, Lowe TM, Mathews MB, O’Reilly D, et al. A guide to naming human non-coding RNA genes. EMBO J. 2020;39:e103777.

213. Seal RL, Tweedie S, Bruford EA. A standardised nomenclature for long non-coding RNAs. IUBMB Life. 2023;75:380–9.

214. HGNC. MicroRNA non-coding host genes | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/1688. Accessed 24 Aug 2023.

215. HGNC. Small nucleolar RNA non-coding host genes | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/1839. Accessed 24 Aug 2023.

216. HGNC. Divergent transcripts | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/1990. Accessed 24 Aug 2023.

217. HGNC. Intronic transcripts | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/1989. Accessed 24 Aug 2023.

218. HGNC. Overlapping transcripts | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/1988. Accessed 24 Aug 2023.

219. HGNC. Antisense RNAs | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/1987. Accessed 24 Aug 2023.

220. HGNC. Long intergenic non-protein coding RNAs | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/1986. Accessed 24 Aug 2023.

221. HGNC. Long non-coding RNAs with FAM root symbol | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/1993. Accessed 24 Aug 2023.

222. HGNC. Long non-coding RNAs | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/788. Accessed 24 Aug 2023.

223. Xiyuan L, Dechao B, Liang S, Yang W, Shuangsang F, Hui L, et al. Using the NONCODE Database Resource. Current Protocols in Bioinformatics. 2017;58:12.16.1-12.16.19.

224. Volders P-J, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Research. 2019;47:D135–9.

225. Fitch WM. Distinguishing Homologous from Analogous Proteins. Systematic Biology. 1970;19:99–113.

226. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLOS Computational Biology. 2012;8:e1002514.

227. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309–38.

228. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. PLOS Computational Biology. 2011;7:e1002073.

229. Ingram CJE, Mulcare CA, Itan Y, Thomas MG, Swallow DM. Lactose digestion and the evolutionary genetics of lactase persistence. Hum Genet. 2009;124:579–91.

230. Wells JCK, Pomeroy E, Stock JT. Evolution of Lactase Persistence: Turbo-Charging Adaptation in Growth Under the Selective Pressure of Maternal Mortality? Front Physiol. 2021;12:696516.

231. de Rosa R, Grenier JK, Andreeva T, Cook CE, Adoutte A, Akam M, et al. Hox genes in brachiopods and priapulids and protostome evolution. Nature. 1999;399:772–6.

232. Krumlauf R. Hox genes in vertebrate development. Cell. 1994;78:191–201.

233. Gaunt SJ. Evolutionary Developmental Biology: Homologous Regulatory Genes and Processes. In: John Wiley & Sons, Ltd, editor. eLS. 1st edition. Wiley; 2001.

234. Zahn-Zabal M, Dessimoz C, Glover NM. Identifying orthologs with OMA: A primer. F1000Res. 2020;9:27.

235. EMBL-EBI. Ensembl genome browser - Orthologues View. 2023. https://www.ensembl.org/Help/View?id=135. Accessed 25 Aug 2023.

236. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656–64.

237. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.

238. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21:487–93.

239. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.

240. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.

241. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

242. Hernández-Salmerón JE, Moreno-Hagelsieb G. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. BMC Genomics. 2020;21:741.

243. Persson E, Sonnhammer ELL. InParanoiDB 9: Ortholog Groups for Protein Domains and Full-Length Proteins. Journal of Molecular Biology. 2023;435:168001.

244. Train C-M, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C. Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. Bioinformatics. 2017;33:i75–82.

245. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology. 2019;20:238.

246. Pevzner P, Tesler G. Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes. Genome Res. 2003;13:37–45.

247. Ng M-P, Vergara IA, Frech C, Chen Q, Zeng X, Pei J, et al. OrthoClusterDB: an online platform for synteny blocks. BMC Bioinformatics. 2009;10:192.

248. Zeng X, Nesbitt MJ, Pei J, Wang K, Vergara IA, Chen N. OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics. In: Proceedings of the 11th international conference on Extending database technology: Advances in database technology. New York, NY, USA: Association for Computing Machinery; 2008. p. 656–67.

249. EMBL-EBI. Orthology quality-controls. 2023. https://www.ensembl.org/info/genome/compara/Ortholog_qc_manual.html. Accessed 25 Aug 2023.

250. Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 2006;34 Database issue:D572–80.

251. van der Heijden RT, Snel B, van Noort V, Huynen MA. Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics. 2007;8:83.

252. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 2019;47:D309–14.

253. Altenhoff AM, Train C-M, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, et al. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids Research. 2021;49:D373–9.

254. OMA. OMA browser. 2023. https://omabrowser.org./oma/home/. Accessed 25 Aug 2023.

255. Computational Biology group - EMBL, Heidelberg. EggNOG 5.0.0. 2023. http://eggnog5.embl.de/#/app/home. Accessed 25 Aug 2023.

256. Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, et al. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Research. 2023;51:D445–51.

257. OrthoDB. OrthoDB. https://www.orthodb.org/. Accessed 25 Aug 2023.

258. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl comparative genomics resources. Database (Oxford). 2016;2016:bav096.

259. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart – biological queries made easy. BMC Genomics. 2009;10:22.

260. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496:498–503.

261. Carroll SB. Evolution at Two Levels: On Genes and Form. PLOS Biology. 2005;3:e245.

262. Barrett RDH, Hoekstra HE. Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet. 2011;12:767–80.

263. Gallegos JE, Rose AB. Intron DNA Sequences Can Be More Important Than the Proximal Promoter in Determining the Site of Transcript Initiation. Plant Cell. 2017;29:843–53.

264. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11:345–55.

265. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science. 2012;338:1587–93.

266. Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, et al. Identification and Characterization of the Potential Promoter Regions of 1031 Kinds of Human Genes. Genome Res. 2001;11:677–84.

267. Smale ST, Kadonaga JT. The RNA Polymerase II Core Promoter. Annual Review of Biochemistry. 2003;72:449–79.

268. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8:206–16.

269. Mayr C. Evolution and Biological Roles of Alternative 3'UTRs. Trends Cell Biol. 2016;26:227–37.

270. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 2003;4:865–75.

271. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet. 2005;21:673–82.

272. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24:1963–76.

273. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 2012;46:21–42.

274. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Reports. 2015;11:1110–22.

275. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465:1033–8.

276. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505:635–40.

277. Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, et al. Origin of Primate Orphan Genes: A Comparative Genomics Approach. Molecular Biology and Evolution. 2009;26:603–12.

278. Johnsson P, Ackley A, Vidarsdottir L, Lui W-O, Corcoran M, Grandér D, et al. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. Nat Struct Mol Biol. 2013;20:440–6.

279. Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, et al. A Dual Origin of the Xist Gene from a Protein-Coding Gene and a Set of Transposable Elements. PLoS One. 2008;3:e2521.

280. Shevchenko A, Zakharova I, Zakian S. The Evolutionary Pathway of X Chromosome Inactivation in Mammals. Acta naturae. 2013;5:40–53.

281. Duret L, Chureau C, Samain S, Weissenbach J, Avner P. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science. 2006;312:1653–5.

282. He S, Liu S, Zhu H. The sequence, structure and evolutionary features of HOTAIR in mammals. BMC Evolutionary Biology. 2011;11:102.

283. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat Rev Genet. 2016;17:601–14.

284. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. PLOS Genetics. 2012;8:e1002841.

285. Washietl S, Findeiß S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, et al. RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. RNA. 2011;17:578–94.

286. Ross CJ, Rom A, Spinrad A, Gelbard-Solodkin D, Degani N, Ulitsky I. Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. Genome Biology. 2021;22:29.

287. Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet. 2014;30:439–52.

288. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biology. 2012;13:R107.

289. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.

290. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.

291. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133:523–36.

292. Marguerat S, Bähler J. RNA-seq: from technology to biology. Cell Mol Life Sci. 2010;67:569–79.

293. Cooper DA, Cortés-López M, Miura P. Genome-Wide circRNA Profiling from RNA-seq Data. In: Dieterich C, Papantonis A, editors. Circular RNAs: Methods and Protocols. New York, NY: Springer; 2018. p. 27–41.

294. O'Neil D, Glowatz H, Schlumpberger M. Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity. Current Protocols in Molecular Biology. 2013;103:4.19.1-4.19.8.

295. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;2015:951–69.

296. Eisenstein M. Illumina faces short-read rivals. Nature Biotechnology. 2023;41:3–5.

297. Franklin Carpenter. The Top 5 Genome Sequencing Companies by Revenue. BioSpace. 2022. https://www.biospace.com/article/top-10-gene-sequencing-companies-by-revenue/. Accessed 25 Aug 2023.

298. Sullivan M. Sequencing 101: long-read sequencing. PacBio. 2023. https://www.pacb.com/blog/long-read-sequencing/. Accessed 25 Aug 2023.

299. DNA sequencing | Oxford Nanopore Technologies. https://nanoporetech.com/applications/dna-nanopore-sequencing. Accessed 25 Aug 2023.

300. Morisse P, Marchet C, Limasset A, Lecroq T, Lefebvre A. Scalable long read self-correction and assembly polishing with multiple sequence alignment. Sci Rep. 2021;11:761.

301. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2010;38:1767–71.

302. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12:671–82.

303. Geniza M, Jaiswal P. Tools for building de novo transcriptome assembly. Current Plant Biology. 2017;11–12:41–5.

304. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to de novo transcriptome assembly and annotation. Briefings in Bioinformatics. 2022;23:bbab563.

305. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 2016;26:1134–44.

306. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

307. Dobin A. STAR 2.7.11a. 2023.

308. Zhang Y, Park C, Bennett C, Thornton M, Kim D. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. Genome Res. 2021;31:1290–5.

309. HISAT2. 2023.

310. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. 2013;14:R36.

311. Kim D. infphilo/tophat. 2023.

312. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

313. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37:4572–4.

314. Li H. lh3/minimap2. 2023.

315. LoTempio J, Délot E, Vilain E. Benchmarking long-read genome sequence alignment tools for human genomics applications. 2021;:2021.07.09.451840.

316. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.

317. Rescheneder P. philres/ngmlr. 2023.

318. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun. 2016;7:11307.

319. Sovic I. isovic/graphmap. 2023.

320. Liu B, Gao Y, Wang Y. LAMSA: fast split read alignment with long approximate matches. Bioinformatics. 2017;33:192–201.

321. LAMSA. 2020.

322. FR-AgENCODE · project overview. https://www.fragencode.org/overview.html. Accessed 25 Aug 2023.

323. Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, et al. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biology. 2019;17:108.

324. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLOS ONE. 2017;12:e0190152.

325. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Briefings in Bioinformatics. 2019;20:2044–54.

326. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;11:R106.

327. thelovelab/DESeq2. 2023.

328. Chen Y, Lun AT, McCarthy DJ, Ritchie ME, Phipson B, Hu Y, et al. edgeR: Empirical Analysis of Digital Gene Expression Data in R. 2023.

329. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

330. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. Bioinform Biol Insights. 2015;9s1:BBI.S28991.

331. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

332. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18:1509–17.

333. Wang J, Dean DC, Hornicek FJ, Shi H, Duan Z. RNA sequencing (RNA-Seq) and its application in ovarian cancer. Gynecologic Oncology. 2019;152:194–201.

334. Liang J, Lv J, Liu Z. Identification of stage-specific biomarkers in lung adenocarcinoma based on RNA-seq data. Tumor Biol. 2015;36:6391–9.

335. Saliba A-E, C Santos S, Vogel J. New RNA-seq approaches for the study of bacterial pathogens. Current Opinion in Microbiology. 2017;35:78–87.

336. Creecy JP, Conway T. Quantitative bacterial transcriptomics with RNA-seq. Current Opinion in Microbiology. 2015;23:133–40.

337. Piskol R, Ramaswami G, Li JB. Reliable Identification of Genomic Variants from RNA-Seq Data. The American Journal of Human Genetics. 2013;93:641–51.

338. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, et al. Identifying RNA editing sites using RNA sequencing data alone. Nat Methods. 2013;10:128–32.

339. Park E, Williams B, Wold BJ, Mortazavi A. RNA editing in the human ENCODE RNA-seq data. Genome Res. 2012;22:1626–33.

340. Mansi L, Tangaro MA, Lo Giudice C, Flati T, Kopel E, Schaffer AA, et al. REDIportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. Nucleic Acids Research. 2021;49:D1012–9.

341. Peng Z, Cheng Y, Tan BC-M, Kang L, Tian Z, Zhu Y, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol. 2012;30:253–60.

342. Khermesh K, D'Erchia AM, Barak M, Annese A, Wachtel C, Levanon EY, et al. Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. RNA. 2016;22:290–302.

343. Gallo A, Vukic D, Michalík D, O'Connell MA, Keegan LP. ADAR RNA editing in human disease; more to it than meets the I. Hum Genet. 2017;136:1265–78.

344. Lagarrigue S, Hormozdiari F, Martin LJ, Lecerf F, Hasin Y, Rau C, et al. Limited RNA Editing in Exons of Mouse Liver and Adipose. Genetics. 2013;193:1107–15.

345. Frésard L, Leroux S, Servin B, Gourichon D, Dehais P, Cristobal MS, et al. Transcriptome-wide investigation of genomic imprinting in chicken. Nucleic Acids Research. 2014;42:3768–82.

346. Roux P-F, Frésard L, Boutin M, Leroux S, Klopp C, Djari A, et al. The Extent of mRNA Editing Is Limited in Chicken Liver and Adipose, but Impacted by Tissular Context, Genotype, Age, and Feeding as Exemplified with a Conserved Edited Site in COG3. G3 Genes|Genomes|Genetics. 2016;6:321–35.

347. Shafiei H, Bakhtiarizadeh MR, Salehi A. Large-scale potential RNA editing profiling in different adult chicken tissues. Animal Genetics. 2019;50:460–74.

348. Kleinman CL, Adoue V, Majewski J. RNA editing of protein sequences: A rare event in human transcriptomes. RNA. 2012;18:1586–96.

349. Tan MH, Li Q, Shanmugam R, Piskol R, Kohler J, Young AN, et al. Dynamic landscape and regulation of RNA editing in mammals. Nature. 2017;550:249–54.

350. Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. 2017;45:D750–7.

351. Pirinen M, Lappalainen T, Zaitlen NA, GTEx Consortium, Dermitzakis ET, Donnelly P, et al. Assessing allele-specific expression across multiple tissues from RNA-seq read data. Bioinformatics. 2015;31:2497–504.

352. Lagarrigue S, Martin L, Hormozdiari F, Roux P-F, Pan C, van Nas A, et al. Analysis of Allele-Specific Expression in Mouse Liver by RNA-Seq: A Comparison With Cis-eQTL Identified Using Genetic Linkage. Genetics. 2013;195:1157–66.

353. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct. 2009;4:14.

354. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131:281–5.

355. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

356. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

357. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.

358. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010;11:R25.

359. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:550.

360. Abbas-Aghababazadeh F, Li Q, Fridley BL. Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. PLoS One. 2018;13:e0206312.

361. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.

362. StatQuest: DESeq2, part 1, Library Normalization. 2017.

363. RPKM, FPKM and TPM, Clearly Explained!!! 2015.

364. StatQuest: edgeR and DESeq2, part 2 - Independent Filtering. 2017.

365. StatQuest : Principal Component Analysis (PCA) (étape par étape). 2018.

366. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun. 2017;8:16027.

367. Byrne A, Cole C, Volden R, Vollmers C. Realizing the potential of full-length transcriptome sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences. 2019;374:20190097.

368. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biology. 2020;21:30.

369. Hu Y, Fang L, Chen X, Zhong JF, Li M, Wang K. LIQA: long-read isoform quantification and analysis. Genome Biology. 2021;22:182.

370. Gao Y, Wang F, Wang R, Kutschera E, Xu Y, Xie S, et al. ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. Science Advances. 2023;9:eabq5072.

371. Gleeson J, Leger A, Prawer YDJ, Lane TA, Harrison PJ, Haerty W, et al. Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. Nucleic Acids Research. 2022;50:e19.

372. Guan D, Halstead MM, Islas-Trejo AD, Goszczynski DE, Cheng HH, Ross PJ, et al. Prediction of transcript isoforms in 19 chicken tissues by Oxford Nanopore long-read sequencing. Frontiers in Genetics. 2022;13.

373. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA. 2020;26:903–9.

374. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;17:13.

375. Maza E. In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. Frontiers in Genetics. 2016;7.

376. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics. 2013;14:671–83.

377. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18:205–14.

378. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21:650–9.

379. Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, Schütz F, et al. Mechanisms and Evolutionary Patterns of Mammalian and Avian Dosage Compensation. PLoS Biol. 2012;10:e1001328.

380. Schug J, Schuller W-P, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ. Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biology. 2005;6:R33.

381. Huminiecki L, Lloyd AT, Wolfe KH. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. BMC Genomics. 2003;4:31.

382. Vandenbon A, Nakai K. Modeling tissue-specific structural patterns in human and mouse promoters. Nucleic Acids Res. 2010;38:17–25.

383. Palmer D, Fabris F, Doherty A, Freitas AA, Magalhães JP de. Ageing transcriptome meta-analysis reveals similarities and differences between key mammalian tissues. Aging. 2021;13:3313–41.

384. Lüleci HB, Yılmaz A. Robust and rigorous identification of tissue-specific genes by statistically extending tau score. BioData Mining. 2022;15:31.

385. Lizio M, Deviatiiarov R, Nagai H, Galan L, Arner E, Itoh M, et al. Systematic analysis of transcription start sites in avian development. PLOS Biology. 2017;15:e2002887.

386. Zhao L, Wit J, Svetec N, Begun DJ. Parallel Gene Expression Differences between Low and High Latitude Populations of Drosophila melanogaster and D. simulans. PLOS Genetics. 2015;11:e1005184.

387. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of The Royal Society of Edinburgh. 1919;52:399–433.

388. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

389. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.

390. EMBL-EBI. Index of /pub/release-110/variation/gvf/homo_sapiens. 2023. https://ftp.ensembl.org/pub/release-110/variation/gvf/homo_sapiens/. Accessed 23 Oct 2023.

391. EMBL-EBI. Index of /pub/release-110/variation/gvf/gallus_gallus. 2023. https://ftp.ensembl.org/pub/release-110/variation/gvf/gallus_gallus/. Accessed 27 Sep 2023.

392. Fu W, Wang R, Xu N, Wang J, Li R, Asadollahpour Nanaei H, et al. Galbase: a comprehensive repository for integrating chicken multi-omics data. BMC Genomics. 2022;23:364.

393. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology. 2012;8:e1002822.

394. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518:197–206.

395. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46:1173–86.

396. Elks CE, den Hoed M, Zhao JH, Sharp SJ, Wareham NJ, Loos RJF, et al. Variability in the heritability of body mass index: a systematic review and meta-regression. Front Endocrinol (Lausanne). 2012;3:29.

397. Matthews LJ, Turkheimer E. Three Legs of the Missing Heritability Problem. Stud Hist Philos Sci. 2022;93:183–91.

398. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.

399. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–9.

400. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169:1177–86.

401. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94:559–73.

402. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.

403. Gross DS, Garrard WT. Nuclease Hypersensitive Sites in Chromatin. Annual Review of Biochemistry. 1988;57:159–97.

404. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132:311–22.

405. Guan D, Bai Z, Zhu X, Zhong C, Hou Y, Consortium TC, et al. The ChickenGTEx pilot analysis: a reference of regulatory variants across 28 chicken tissues. 2023;:2023.06.27.546670.

406. The CattleGTEx atlas reveals regulatory mechanisms underlying complex traits. Nat Genet. 2022;54:1273–4.

407. Dixon JR, Gorkin DU, Ren B. Chromatin Domains: The Unit of Chromosome Organization. Mol Cell. 2016;62:668–80.

408. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5:eaaw1668.

409. Fishman V, Battulin N, Nuriddinov M, Maslova A, Zlotina A, Strunov A, et al. 3D organization of chicken genome demonstrates evolutionary conservation of topologically associated domains and highlights unique architecture of erythrocytes' chromatin. Nucleic Acids Res. 2019;47:648–65.

410. Teumer A, Ernst FD, Wiechert A, Uhr K, Nauck M, Petersmann A, et al. Comparison of genotyping using pooled DNA samples (allelotyping) and individual genotyping using the affymetrix genome-wide human SNP array 6.0. BMC Genomics. 2013;14:506.

411. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013;14:59.

412. Trotter C, Kim H, Farage G, Prins P, Williams RW, Broman KW, et al. Speeding up eQTL scans in the BXD population using GPUs. G3 (Bethesda). 2021;11:jkab254.

413. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.

414. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 2016;32:1479–85.

415. Taylor-Weiner A, Aguet F, Haradhvala NJ, Gosai S, Anand S, Kim J, et al. Scaling computational genomics to millions of individuals with GPUs. Genome Biology. 2019;20:228.

416. Ma J, Chen S, Hao L, Sheng W, Chen W, Ma X, et al. Hypermethylation-mediated down-regulation of lncRNA TBX5-AS1:2 in Tetralogy of Fallot inhibits cell proliferation by reducing TBX5 expression. Journal of Cellular and Molecular Medicine. 2020;24:6472–84.

417. Hu Z-L, Park CA, Reecy JM. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Research. 2022;50:D956–61.

418. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLOS Genetics. 2014;10:e1004383.

419. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99:1245–60.

420. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLOS Genetics. 2017;13:e1006646.

421. Beesley J, Sivakumaran H, Marjaneh MM, Shi W, Hillman KM, Kaufmann S, et al. eQTL Colocalization Analyses Identify NTN4 as a Candidate Breast Cancer Risk Gene. The American Journal of Human Genetics. 2020;107:778–87.

422. EMBL-EBI. Homo_sapiens - Ensembl genome browser 109. 2023. https://feb2023.archive.ensembl.org/Homo_sapiens/Info/Annotation. Accessed 2 Oct 2023.

423. EMBL-EBI. Mus_musculus - Ensembl genome browser 109. 2023. https://feb2023.archive.ensembl.org/Mus_musculus/Info/Annotation. Accessed 2 Oct 2023.

424. Jiang S, Cheng S-J, Ren L-C, Wang Q, Kang Y-J, Ding Y, et al. An expanded landscape of human long noncoding RNA. Nucleic Acids Res. 2019;47:7842–56.

425. Marx V. Method of the year: long-read sequencing. Nat Methods. 2023;20:6–11.

426. Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet. 2017;49:1731–40.

427. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, et al. Control of pancreas and liver gene expression by HNF transcription factors. Science. 2004;303:1378–81.

428. Dhir A, Dhir S, Proudfoot NJ, Jopling CL. Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. Nat Struct Mol Biol. 2015;22:319–27.

429. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604:310–5.

430. Xu Q, Song Z, Zhu C, Tao C, Kang L, Liu W, et al. Systematic comparison of lncRNAs with protein coding mRNAs in population expression and their response to environmental change. BMC Plant Biol. 2017;17:42.

431. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat Commun. 2019;10:3359.

432. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. 2013;3:2179–90.

433. Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. Developmental dynamics of lncRNAs across mammalian organs and species. Nature. 2019;571:510–4.

434. Hezroni H, Ben-Tov Perry R, Meir Z, Housman G, Lubelsky Y, Ulitsky I. A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. Genome Biology. 2017;18:162.

435. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. Science. 2020;369:eaba3066.

436. Rinn JL, Snyder M. Sexual dimorphism in mammalian gene expression. Trends Genet. 2005;21:298–305.

437. García-Calzón S, Perfilyev A, de Mello VD, Pihlajamäki J, Ling C. Sex Differences in the Methylome and Transcriptome of the Human Liver and Circulating HDL-Cholesterol Levels. J Clin Endocrinol Metab. 2018;103:4395–408.

438. Liu B, Shyr Y, Cai J, Liu Q. Interplay between miRNAs and host genes and their role in cancer. Brief Funct Genomics. 2018;18:255–66.

439. Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 2005;11:241–7.

440. Dohi O, Yasui K, Gen Y, Takada H, Endo M, Tsuji K, et al. Epigenetic silencing of miR-335 and its host gene MEST in hepatocellular carcinoma. Int J Oncol. 2013;42:411–8.

441. Cai Y, Yu X, Hu S, Yu J. A brief review on the mechanisms of miRNA regulation. Genomics Proteomics Bioinformatics. 2009;7:147–54.

442. Kim Y-K, Kim VN. Processing of intronic microRNAs. EMBO J. 2007;26:775–83.

443. Ozsolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, et al. Chromatin structure analyses identify miRNA promoters. Genes Dev. 2008;22:3172–83.

444. Kern C, Wang Y, Chitwood J, Korf I, Delany M, Cheng H, et al. Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. BMC Genomics. 2018;19:684.

445. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458:223–7.

446. NCBI-RefSeq. bGalGal1.mat.broiler.GRCg7b - Genome - Assembly - NCBI. NCBI. 2021. https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_016699485.2/. Accessed 14 Sep 2023.

447. NCBI-RefSeq. bGalGal1.mat.broiler.GRCg7b - Genome - Annotation - NCBI v106. 2022. https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/699/485/GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b/. Accessed 14 Sep 2023.

448. EMBL EBI Ensembl/GENCODE. bGalGal1.mat.broiler.GRCg7b - Genome - Annotation - Ensembl v107. 2022. https://ftp.ensembl.org/pub/release-107/gtf/gallus_gallus/. Accessed 14 Sep 2023.

449. NCBI-RefSeq. Coordinate remapping service: NCBI. 2022. https://www.ncbi.nlm.nih.gov/genome/tools/remap. Accessed 14 Sep 2023.

450. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

451. Patel H, Ewels P, Peltzer A, Hammarén R, Botvinnik O, Sturm G, et al. nf-core/rnaseq: nf-core/rnaseq v3.8.1 - Plastered Magnesium Mongoose. 2022.

452. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38:276–8.

453. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. 2008;25:1–18.

454. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

455. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014;42:e91.

456. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57:289–300.

457. FarmGTEx - Farm Animal Genotype-Tissue Expression. FarmGTEx. https://www.farmgtex.org/. Accessed 13 Sep 2023.

458. Chapman DL, Garvey N, Hancock S, Alexiou M, Agulnik SI, Gibson-Brown JJ, et al. Expression of the T-box family genes, Tbx1-Tbx5, during early mouse development. Dev Dyn. 1996;206:379–90.

459. Hatcher C, Kim M, Maha C, Goldstein M, Wong B, Mikawa T, et al. TBX5 transcription factor regulates cell proliferation during cardiogenesis. Developmental biology. 2001;230:177–88.

460. Hori Y, Tanimoto Y, Takahashi S, Furukawa T, Koshiba-Takeuchi K, Takeuchi JK. Important cardiac transcription factor genes are accompanied by bidirectional long non-coding RNAs. BMC Genomics. 2018;19:967.

461. Douaud M, Feve K, Pituello F, Gourichon D, Boitard S, Leguern E, et al. Epilepsy Caused by an Abnormal Alternative Splicing with Dosage Effect of the SV2A Gene in a Chicken Model. PLoS One. 2011;6:e26932.

462. Hamilton JA, Hillard CJ, Spector AA, Watkins PA. Brain uptake and utilization of fatty acids, lipids and lipoproteins: application to neurological disorders. J Mol Neurosci. 2007;33:2–11.

463. Montgomery SB, Bernstein JA, Wheeler MT. Toward transcriptomics as a primary tool for rare disease investigation. Cold Spring Harb Mol Case Stud. 2022;8:a006198.

464. Frankish A, Uszczynska B, Ritchie GR, Gonzalez JM, Pervouchine D, Petryszak R, et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 2015;16:S2.

465. Wu P-Y, Phan JH, Wang MD. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinformatics. 2013;14:S8.

466. Chisanga D, Liao Y, Shi W. Impact of gene annotation choice on the quantification of RNA-seq data. BMC Bioinformatics. 2022;23:107.

467. Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;16:97.

468. Wu P-Y, Phan JH, Wang MD. The Effect of Human Genome Annotation Complexity on RNA-Seq Gene Expression Quantification. IEEE Int Conf Bioinform Biomed Workshops. 2012;2012:712–7.

469. Hamaguchi Y, Zeng C, Hamada M. Impact of human gene annotations on RNA-seq differential expression analysis. BMC Genomics. 2021;22:730.

470. Chen G, Wang C, Shi L, Qu X, Chen J, Yang J, et al. Incorporating the human gene annotations in different databases significantly improved transcriptomic and genetic analyses. RNA. 2013;19:479–89.

471. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing Reference Genome Assemblies. PLOS Biology. 2011;9:e1001091.

472. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics. 2017;109:83–90.

473. Lansdon LA, Cadieux-Dion M, Yoo B, Miller N, Cohen ASA, Zellmer L, et al. Factors Affecting Migration to GRCh38 in Laboratories Performing Clinical Next-Generation Sequencing. J Mol Diagn. 2021;23:651–7.

474. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53.

475. Frésard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med. 2019;25:911–9.

476. Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. 1990.

477. Park J, Yhim H-Y, Kang KP, Bae TW, Cho YG. Copy number variation analysis using next-generation sequencing identifies the CFHR3/CFHR1 deletion in atypical hemolytic uremic syndrome: a case report. Hematology. 2022;27:603–8.

478. Zipfel PF, Edey M, Heinen S, Józsi M, Richter H, Misselwitz J, et al. Deletion of complement factor H-related genes CFHR1 and CFHR3 is associated with atypical hemolytic uremic syndrome. PLoS Genet. 2007;3:e41.

479. Hamza A, El Sissy C, Yousfi N, Martins P, Rafat C, Masliah-Planchon J, et al. The absence of CFHR3 and CFHR1 genes from the T2T-CHM13 assembly can limit the molecular diagnosis of complement-related diseases. European journal of human genetics : EJHG. 2023;31.

480. Hansen J, Snow C, Tuttle E, Ghoneim DH, Yang C-S, Spencer A, et al. De Novo Mutations in SIK1 Cause a Spectrum of Developmental Epilepsies. Am J Hum Genet. 2015;96:682–90.

481. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27:849–64.

482. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research. 2015;43:D789–98.

483. Ohba C, Kato M, Takahashi S, Lerman-Sagie T, Lev D, Terashima H, et al. Early onset epileptic encephalopathy caused by de novo SCN8A mutations. Epilepsia. 2014;55:994–1000.

484. Tang D, Zhao X, Zhang L, Wang C. Comprehensive analysis of pseudogene HSPB1P1 and its potential roles in hepatocellular carcinoma. J Cell Physiol. 2020;235:6515–27.

485. Takenouchi T, Kodo K, Yamazaki F, Nakatomi H, Kosaki K. Progressive cerebral and coronary aneurysms in the original two patients with Kosaki overgrowth syndrome. American Journal of Medical Genetics Part A. 2021;185:999–1003.

486. Takenouchi T, Yamaguchi Y, Tanikawa A, Kosaki R, Okano H, Kosaki K. Novel Overgrowth Syndrome Phenotype Due to Recurrent De Novo PDGFRB Mutation. The Journal of Pediatrics. 2015;166:483–6.

487. Kirk JM, Kim SO, Inoue K, Smola MJ, Lee DM, Schertzer MD, et al. Functional classification of long non-coding RNAs by kmer content. Nat Genet. 2018;50:1474–82.

488. Bryzghalov O, Szcześniak MW, Makałowska I. SyntDB: defining orthologues of human long noncoding RNAs across primates. Nucleic Acids Research. 2020;48:D238–45.

489. EMBL-EBI. Details on a Compara analysis - 63 amniota vertebrates. 2023. https://www.ensembl.org/info/genome/compara/mlss.html?mlss=2041. Accessed 28 Oct 2023.

490. Gondret F, Vincent A, Houée-Bigot M, Siegel A, Lagarrigue S, Causeur D, et al. A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics. 2017;18:244.

491. Savary C, Kim A, Lespagnol A, Gandemer V, Pellier I, Andrieu C, et al. Depicting the genetic architecture of pediatric cancers through an integrative gene network approach. Sci Rep. 2020;10:1224.

492. Jehl F, Désert C, Klopp C, Brenet M, Rau A, Leroux S, et al. Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. BMC Genomics. 2019;20:1033.

493. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24:14–24.

494. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.

495. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.

496. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40:1413–5.

497. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.

498. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, et al. Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. PLOS ONE. 2013;8:e58815.

499. Tang X, Baheti S, Shameer K, Thompson KJ, Wills Q, Niu N, et al. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. Nucleic Acids Research. 2014;42:e172.

500. Wang C, Davila JI, Baheti S, Bhagwate AV, Wang X, Kocher J-PA, et al. RVboost: RNA-seq variants prioritization using a boosting method. Bioinformatics. 2014;30:3414–6.

501. Wolfien M, Rimmbach C, Schmitz U, Jung JJ, Krebs S, Steinhoff G, et al. TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. BMC Bioinformatics. 2016;17:21.

502. Oikkonen L, Lise S. Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. Wellcome Open Res. 2017;2:6.

503. Cornwell M, Vangala M, Taing L, Herbert Z, Köster J, Li B, et al. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC Bioinformatics. 2018;19:135.

504. Adetunji MO, Lamont SJ, Abasht B, Schmidt CJ. Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. PLOS ONE. 2019;14:e0216838.

505. Zhao Z, Fu Y-X, Hewett-Emmett D, Boerwinkle E. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene. 2003;312:207–13.

506. Guo Y, Zhao S, Sheng Q, Samuels DC, Shyr Y. The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. BMC Genomics. 2017;18:690.

507. Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biology. 2014;12:42.

508. EMBL-EBI. Index of pub/release-94/variation/gvf/gallus_gallus/. 2018. ftp://ftp.ensembl.org/pub/release-94/variation/gvf/gallus_gallus/. Accessed 23 Oct 2023.

509. Zhuo Z, Lamont SJ, Abasht B. RNA-Seq Analyses Identify Frequent Allele Specific Expression and No Evidence of Genomic Imprinting in Specific Embryonic Tissues of Chicken. Sci Rep. 2017;7:11944.

510. Bordas A, Tixier-Boichard M, Merat P. Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. British Poultry Science. 1992;33:741–54.

511. Leclercq B, Simon J. Selecting broilers for low or high abdominal fat : observations on the hens during the breeding period. Annales de zootechnie. 1982;31:161–70.

512. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.

513. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 2013;43:11.10.1-11.10.33.

514. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

515. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

516. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

517. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17:122.

518. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–4.

519. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28:3326–8.

520. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nat Commun. 2016;7:12817.

521. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010;38:e164.

522. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012;6:80–92.

523. Wei L, Liu LT, Conroy JR, Hu Q, Conroy JM, Morrison CD, et al. MAC: identifying and correcting annotation for multi-nucleotide variations. BMC Genomics. 2015;16:569.

524. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Research. 2016;44:e108.

525. Cheng S-J, Shi F-Y, Liu H, Ding Y, Jiang S, Liang N, et al. Accurately annotate compound effects of genetic variants using a context-sensitive framework. Nucleic Acids Research. 2017;45:e82.

526. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017;33:2037–9.

527. Khan W, Varma Saripella G, Ludwig T, Cuppens T, Thibord F, Génin E, et al. MACARON: a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data. Bioinformatics. 2018;34:3396–8.

528. Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat Commun. 2020;11:2539.

529. Jehl F, Degalez F, Bernard M, Lecerf F, Lagoutte L, Désert C, et al. RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Front Genet. 2021;12:655707.

530. Gimeno RE, Hirsch DJ, Punreddy S, Sun Y, Ortegon AM, Wu H, et al. Targeted deletion of fatty acid transport protein-4 results in early embryonic lethality. J Biol Chem. 2003;278:49512–6.

531. Herrmann T, van der Hoeven F, Grone H-J, Stewart AF, Langbein L, Kaiser I, et al. Mice with targeted disruption of the fatty acid transport protein 4 (Fatp 4, Slc27a4) gene show features of lethal restrictive dermopathy. J Cell Biol. 2003;161:1105–15.

532. Moulson CL, Martin DR, Lugus JJ, Schaffer JE, Lind AC, Miner JH. Cloning of wrinkle-free, a previously uncharacterized mouse mutation, reveals crucial roles for fatty acid transport protein 4 in skin and hair development. Proc Natl Acad Sci U S A. 2003;100:5274–9.

533. Lin M-H, Chang K-W, Lin S-C, Miner JH. Epidermal hyperproliferation in mice lacking fatty acid transport protein 4 (FATP4) involves ectopic EGF receptor and STAT3 signaling. Dev Biol. 2010;344:707–19.

534. Tao J, Koster MI, Harrison W, Moran JL, Beier DR, Roop DR, et al. A Spontaneous Fatp4/Scl27a4 Splice Site Mutation in a New Murine Model for Congenital Ichthyosis. PLOS ONE. 2012;7:e50634.

535. Hirsch D, Stahl A, Lodish HF. A family of fatty acid transporters conserved from mycobacterium to man. Proc Natl Acad Sci U S A. 1998;95:8625–9.

536. Schaffer JE. Fatty acid transport: the roads taken. Am J Physiol Endocrinol Metab. 2002;282:E239-246.

537. ANR - Agence Nationale de la Recherche. Analyse du contrôle génétique de différents caractères sur de longues carrières de production, chez la poule pondeuse. Agence nationale de la recherche. 2021. https://anr.fr/Projet-ANR-20-CE20-0029. Accessed 28 Oct 2023.

538. GEroNIMO. GEroNIMO - Genome and Epigenome eNabled breedIng in MOnogastrics - A Horizon 2020 Project. GEroNIMO. 2021. https://www.geronimo-h2020.eu. Accessed 28 Oct 2023.

539. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81:559–75.

540. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.

541. GitHub - nf-core/rnavar at 1.0.0. GitHub. https://github.com/nf-core/rnavar. Accessed 29 Oct 2023.

542. FAO. FAOSTAT. 2023. https://www.fao.org/faostat/en/#data/QCL/visualize. Accessed 4 Oct 2023.

543. Burt DW. Emergence of the chicken as a model organism: implications for agriculture and biology. Poult Sci. 2007;86:1460–71.

544. Beacon TH, Davie JR. The chicken model organism for epigenomic research. Genome. 2021;64:476–89.

545. Garcia P, Wang Y, Viallet J, Macek Jilkova Z. The Chicken Embryo Model: A Novel and Relevant Model for Immune-Based Studies. Front Immunol. 2021;12:791081.

546. Wright D, Rubin C-J, Martinez Barrio A, Schütz K, Kerje S, Brändström H, et al. The genetic architecture of domestication in the chicken: effects of pleiotropy and linkage. Mol Ecol. 2010;19:5140–56.

547. Flores-Santin J, Burggren WW. Beyond the Chicken: Alternative Avian Models for Developmental Physiological Research. Front Physiol. 2021;12:712633.

548. Brown WRA, Hubbard SJ, Tickle C, Wilson SA. The chicken as a model for large-scale analysis of vertebrate gene function. Nat Rev Genet. 2003;4:87–98.

549. Lillie M, Sheng ZY, Honaker CF, Andersson L, Siegel PB, Carlborg Ö. Genomic signatures of 60 years of bidirectional selection for 8-week body weight in chickens. Poultry Science. 2018;97:781–90.

550. van der Eijk JAJ, Verwoolde MB, de Vries Reilingh G, Jansen CA, Rodenburg TB, Lammers A. Chicken lines divergently selected on feather pecking differ in immune characteristics. Physiology & Behavior. 2019;212:112680.

551. Lawal RA, Martin SH, Vanmechelen K, Vereijken A, Silva P, Al-Atiyat RM, et al. The wild species genome ancestry of domestic chickens. BMC Biology. 2020;18:13.

552. Wang M-S, Thakur M, Peng M-S, Jiang Y, Frantz LAF, Li M, et al. 863 genomes reveal the origin and domestication of chicken. Cell Res. 2020;30:693–701.

553. Wang M-S, Zhang J-J, Guo X, Li M, Meyer R, Ashari H, et al. Large-scale genomic analysis reveals the genetic cost of chicken domestication. BMC Biology. 2021;19:118.

554. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464:587–91.

555. Qanbari S, Rubin C-J, Maqbool K, Weigend S, Weigend A, Geibel J, et al. Genetics of adaptation in modern chicken. PLOS Genetics. 2019;15:e1007989.

556. Zhang C, Lin D, Wang Y, Peng D, Li H, Fei J, et al. Widespread introgression in Chinese indigenous chicken breeds from commercial broiler. Evol Appl. 2019;12:610–21.

557. Wu MY, Low GW, Forcina G, van Grouw H, Lee BPY-H, Oh RRY, et al. Historic and modern genomes unveil a domestic introgression gradient in a wild red junglefowl population. Evolutionary Applications. 2020;13:2300–15.

558. Gering E, Johnsson M, Willis P, Getty T, Wright D. Mixed ancestry and admixture in Kauai's feral chickens: invasion of domestic genes into ancient Red Junglefowl reservoirs. Mol Ecol. 2015;24:2112–24.

559. Johnsson M, Gering E, Willis P, Lopez S, Van Dorp L, Hellenthal G, et al. Feralisation targets different genomic loci to domestication in the chicken. Nat Commun. 2016;7:12950.

560. Gheyas AA, Vallejo-Trujillo A, Kebede A, Lozano-Jaramillo M, Dessie T, Smith J, et al. Integrated Environmental and Genomic Analysis Reveals the Drivers of Local Adaptation in African Indigenous Chickens. Mol Biol Evol. 2021;38:4268–85.

561. Zan Y, Sheng Z, Lillie M, Rönnegård L, Honaker CF, Siegel PB, et al. Artificial Selection Response due to Polygenic Adaptation from a Multilocus, Multiallelic Genetic Architecture. Molecular Biology and Evolution. 2017;34:2678–89.

562. Shi S, Shao D, Yang L, Liang Q, Han W, Xue Q, et al. Whole genome analyses reveal novel genes associated with chicken adaptation to tropical and frigid environments. J Adv Res. 2023;47:13–25.

563. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10:381–91.

564. Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet. 2004;5:202–12.

565. Pan Z, Wang Y, Wang M, Wang Y, Zhu X, Gu S, et al. An atlas of regulatory elements in chicken: A resource for chicken genetics and genomics. Science Advances. 2023;9:eade1204.

566. Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat Commun. 2021;12:1821.

567. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16:197–212.

568. Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet. 2021;53:1290–9.

569. Johnsson M, Jonsson KB, Andersson L, Jensen P, Wright D. Genetic Regulation of Bone Metabolism in the Chicken: Similarities and Differences to Mammalian Systems. PLOS Genetics. 2015;11:e1005250.

570. Höglund A, Strempfl K, Fogelholm J, Wright D, Henriksen R. The genetic regulation of size variation in the transcriptome of the cerebrum in the chicken and its role in domestication and brain size evolution. BMC Genomics. 2020;21:518.

571. Falker-Gieske C, Bennewitz J, Tetens J. Structural variation and eQTL analysis in two experimental populations of chickens divergently selected for feather-pecking behavior. Neurogenetics. 2023;24:29–41.

572. Mott AC, Mott A, Preuß S, Bennewitz J, Tetens J, Falker-Gieske C. eQTL analysis of laying hens divergently selected for feather pecking identifies KLF14 as a potential key regulator for this behavioral disorder. Front Genet. 2022;13:969752.

573. EMBL-EBI. Gallus_gallus - Ensembl genome browser 102. 2020. https://nov2020.archive.ensembl.org/Gallus_gallus/Info/Index?db=core. Accessed 4 Oct 2023.

574. Consortium TF-P, Teng J, Gao Y, Yin H, Bai Z, Liu S, et al. A compendium of genetic regulatory effects across pig tissues. 2022;:2022.11.11.516073.

575. Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W, et al. A multi-tissue atlas of regulatory variants in cattle. Nat Genet. 2022;54:1438–47.

576. Park K-Y, Hwang H-S, Cho K-H, Han K, Nam GE, Kim YH, et al. Body Weight Fluctuation as a Risk Factor for Type 2 Diabetes: Results from a Nationwide Cohort Study. J Clin Med. 2019;8:950.

577. Fox CS, Heard-Costa N, Cupples LA, Dupuis J, Vasan RS, Atwood LD. Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. BMC Med Genet. 2007;8 Suppl 1 Suppl 1:S18.

578. Qi T, Wu Y, Fang H, Zhang F, Liu S, Zeng J, et al. Genetic control of RNA splicing and its distinct role in complex trait variation. Nat Genet. 2022;54:1355–63.

579. Li L, Huang K-L, Gao Y, Cui Y, Wang G, Elrod ND, et al. An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. Nat Genet. 2021;53:994–1005.

580. Bergman M, Ringertz N. Gene expression pattern of chicken erythrocyte nuclei in heterokaryons. J Cell Sci. 1990;97 (Pt 1):167–75.

581. Désert C, Merlot E, Zerjal T, Bed'hom B, Härtle S, Le Cam A, et al. Transcriptomes of whole blood and PBMC in chickens. Comp Biochem Physiol Part D Genomics Proteomics. 2016;20:1–9.

582. de Klein N, Tsai EA, Vochteloo M, Baird D, Huang Y, Chen C-Y, et al. Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. Nat Genet. 2023;55:377–88.

583. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53:1300–10.

584. Zeng B, Bendl J, Kosoy R, Fullard JF, Hoffman GE, Roussos P. Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. Nat Genet. 2022;54:161–9.

585. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nat Rev Genet. 2019;20:135–56.

586. Li M, Sun C, Xu N, Bian P, Tian X, Wang X, et al. De Novo Assembly of 20 Chicken Genomes Reveals the Undetectable Phenomenon for Thousands of Core Genes on Microchromosomes and Subtelomeric Regions. Molecular Biology and Evolution. 2022;39:msac066.

587. Wang K, Hu H, Tian Y, Li J, Scheben A, Zhang C, et al. The Chicken Pan-Genome Reveals Gene Content Variation and a Promoter Region Deletion in IGF2BP1 Affecting Body Size. Molecular Biology and Evolution. 2021;38:5066–81.

588. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.

589. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017;18:323.

590. Thomas S, Underwood JG, Tseng E, Holloway AK, Subcommittee on behalf of the BTBCI. Long-Read Sequencing of Chicken Transcripts and Identification of New Transcript Isoforms. PLOS ONE. 2014;9:e94650.

591. Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, et al. Cell type-specific genetic regulation of gene expression across human tissues. Science. 2020;369:eaaz8528.

592. Postic C, Girard J. Contribution of de novo fatty acid synthesis to hepatic steatosis and insulin resistance: lessons from genetically engineered mice. J Clin Invest. 2008;118:829–38.

593. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44:821–4.

594. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011;88:76–82.

595. Bashinskaya VV, Kulakova OG, Boyko AN, Favorov AV, Favorova OO. A review of genome-wide association studies for multiple sclerosis: classical and hypothesis-driven approaches. Hum Genet. 2015;134:1143–62.

596. Nassir F, Rector RS, Hammoud GM, Ibdah JA. Pathogenesis and Prevention of Hepatic Steatosis. Gastroenterol Hepatol (N Y). 2015;11:167–75.

597. EMBL-EBI. Automatic annotation using RNA-seq data. 2023. https://www.ensembl.org/info/genome/genebuild/rnaseq_annotation.html. Accessed 26 Oct 2023.

598. NCBI-RefSeq. The NCBI Eukaryotic Genome Annotation Pipeline. 2023. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/. Accessed 26 Oct 2023.

599. Thibaud-Nissen F, Souvorov A, Murphy T, DiCuccio M, Kitts P. Eukaryotic Genome Annotation Pipeline. In: The NCBI Handbook [Internet]. 2nd edition. National Center for Biotechnology Information (US); 2013.

600. Vicente-Saez R, Martinez-Fuentes C. Open Science now: A systematic literature review for an integrated definition. Journal of Business Research. 2018;88:428–36.

601. nf-core/genomeannotator. 2023.

602. EMBL-EBI. Index of /pub/release-110/mysql/gallus_gallus_rnaseq_110_7/analysis.txt. 2023. https://ftp.ensembl.org/pub/release-110/mysql/gallus_gallus_rnaseq_110_7/. Accessed 26 Oct 2023.

603. NCBI-RefSeq. Gallus gallus Annotation Report V106. 2022. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Gallus_gallus/106/. Accessed 26 Oct 2023.

604. National Human Genome Research Institute. Developmental Genotype-Tissue Expression (dGTEx). Developmental Genotype-Tissue Expression (dGTEx). 2023. https://www.genome.gov/Funded-Programs-Projects/Developmental-Genotype-Tissue-Expression. Accessed 26 Oct 2023.

605. Ontology Search - OLS. UBERON:0002107. 2023. https://www.ebi.ac.uk/ols4/ontologies/uberon/classes/http%253A%252F%252Fpurl.obolibrary.org%252Fobo%252FUBERON_0002107. Accessed 26 Oct 2023.

606. Kurylo C, Guyomar C, Foissac S, Djebali S. TAGADA: a scalable pipeline to improve genome annotations with RNA-seq data. NAR Genomics and Bioinformatics. 2023;5:lqad089.

607. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.

608. Miga KH, Sullivan BA. Expanding studies of chromosome structure and function in the era of T2T genomics. Human Molecular Genetics. 2021;30:R198–205.

609. Huang Z, Xu Z, Bai H, Huang Y, Kang N, Ding X, et al. Evolutionary analysis of a complete chicken genome. Proceedings of the National Academy of Sciences. 2023;120:e2216641120.

610. Moreno P, Fexova S, George N, Manning JR, Miao Z, Mohammed S, et al. Expression Atlas update: gene and protein expression in multiple species. Nucleic Acids Research. 2022;50:D129–40.

611. Clough E, Barrett T. The Gene Expression Omnibus Database. In: Mathé E, Davis S, editors. Statistical Genomics: Methods and Protocols. New York, NY: Springer; 2016. p. 93–110.

612. EMBL-EBI. Chicken - Experiments < Expression Atlas < EMBL-EBI. 2023. https://www.ebi.ac.uk/gxa/experiments?species=%22gallus%20gallus%22. Accessed 26 Oct 2023.

613. Zhang X, Cui Y, Ding X, Liu S, Han B, Duan X, et al. Analysis of mRNA-lncRNA and mRNA-lncRNA-pathway co-expression networks based on WGCNA in developing pediatric sepsis. Bioengineered. 2021;12:1457–70.

614. Ning C, Ma T, Hu S, Xu Z, Zhang P, Zhao X, et al. Long Non-coding RNA and mRNA Profile of Liver Tissue During Four Developmental Stages in the Chicken. Frontiers in Genetics. 2020;11.

615. Zhai B, Zhao Y, Li H, Li S, Gu J, Zhang H, et al. Weighted gene co-expression network analysis identified hub genes critical to fatty acid composition in Gushi chicken breast muscle. BMC Genomics. 2023;24:594.

616. Zhao Z, Bai J, Wu A, Wang Y, Zhang J, Wang Z, et al. Co-LncRNA: investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. Database (Oxford). 2015;2015:bav082.

617. Cheng L, Wang P, Tian R, Wang S, Guo Q, Luo M, et al. LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. 2019;47 Database issue:D140–4.

618. Maarouf M, Chen B, Chen Y, Wang X, Rai KR, Zhao Z, et al. Identification of lncRNA-155 encoded by MIR155HG as a novel regulator of innate immunity against influenza A virus infection. Cell Microbiol. 2019;21:e13036.

619. de Goede OM, Nachun DC, Ferraro NM, Gloudemans MJ, Rao AS, Smail C, et al. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. Cell. 2021;184:2633-2648.e19.

620. Villegas VE, Zaphiropoulos PG. Neighboring gene regulation by antisense long non-coding RNAs. Int J Mol Sci. 2015;16:3251–66.

621. Beltran M, Puig I, Peña C, García JM, Alvarez AB, Peña R, et al. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes Dev. 2008;22:756–69.

622. Wight M, Werner A. The functions of natural antisense transcripts. Essays Biochem. 2013;54:91–101.

623. Clark WT, Radivojac P. Analysis of protein function and its prediction from amino acid sequence. Proteins: Structure, Function, and Bioinformatics. 2011;79:2086–96.

624. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, et al. HAMAP in 2015: updates to the protein family classification and annotation system. Nucleic Acids Res. 2015;43 Database issue:D1064-1070.

625. Hérault F, Herry F, Varenne A, Burlot T, Picard–Druet D, Recoquillay J, et al. A linkage disequilibrium study in layers and broiler commercial chicken populations. In: 11. World Congress on Genetics Applied to Livestock Production (WCGALP). Auckland, New Zealand; 2018. p. np.

626. Degalez F, Jehl F, Muret K, Bernard M, Lecerf F, Lagoutte L, et al. Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained. Front Genet. 2021;12:659287.

627. Schneider B, Sweeney BA, Bateman A, Cerny J, Zok T, Szachniuk M. When will RNA get its AlphaFold moment? Nucleic Acids Research. 2023;51:9522–32.

628. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun. 2021;12:941.

629. Brotman SM, Raulerson CK, Vadlamudi S, Currin KW, Shen Q, Parsons VA, et al. Subcutaneous adipose tissue splice quantitative trait loci reveal differences in isoform usage associated with cardiometabolic traits. Am J Hum Genet. 2022;109:66–80.

630. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352:600–4.

631. Gene-Switch. Gene-Switch - Homepage. https://www.gene-switch.eu/. Accessed 29 Oct 2023.

632. Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. 2022;:2022.05.07.491045.

633. Bihan-Duval EL, Nadaf J, Berri C, Pitel F, Graulet B, Godet E, et al. Detection of a Cis eQTL Controlling BMCO1 Gene Expression Leads to the Identification of a QTG for Chicken Breast Meat Color. PLOS ONE. 2011;6:e14825.

COLLEGE ECOLOGIE
DOCTORAL GEOSCIENCES
BRETAGNE AGRONOMIE ALIMENTATION

L'INSTITUT agro Rennes Angers

Titre : Régulation des transcrits, phénotypes intermédiaires entre génotype et phénotype – Focus sur les gènes à ARN longs non-codants chez la poule et les phénotypes liés aux lipides

Mots clés : lncRNA, RNAseq, annotation, orthologie, régulation, eQTL

Résumé : La thèse s'articulait autour de l'annotation du génome de la poule et de l'identification de variants génétiques pour aboutir à des GWAS pour des phénotypes macroscopiques ou expressionnels.
Le premier objectif a été d'enrichir l'annotation GRCg7b du génome de la poule en lncRNA, acteurs clés de la régulation de l'expression des gènes, en combinant six projets multi-tissus et les annotations de référence. L'outil GEGA associé permet de visualiser l'expression de ces gènes au travers de 47 tissus et 1400 échantillons. Des travaux complémentaires sur l'impact des assemblages humains sur la détection d'expressions aberrantes impliquées dans des maladies rares ont été menés. Pour finir, un pipeline pour explorer la conservation des lncRNA entre espèces a permis de présumer des liens d'orthologie pour plusieurs centaines d'entre eux.
Le deuxième objectif visait à l'identification de SNP à partir de 700 RNAseq de poules. Leur intérêt a été évalué pour des études d'ASE, d'exploration de la diversité génétique entre populations et de prédiction des impacts fonctionnels. Un pipeline considérant les phases des SNP au sein d'un même codon a été développé pour affiner les prédictions d'impact sur la protéine. En parallèle, des génotypages obtenus à des années différentes et par puces SNP ont été analysés.
Enfin, le troisième objectif a été de combiner les annotations du génome et les SNP identifiés pour détecter des régions associées à la variabilité de l'expression des gènes (eQTL) et de phénotypes plus complexes (QTL) via le projet ChickenGTEx et une population de poules pondeuses commerciales.

Title: Transcript regulation, intermediate phenotypes between genotype and phenotype – Focus on long non-coding RNA genes in chickens and lipid-related phenotypes

Keywords: lncRNA, RNAseq, annotation, orthology, regulation, eQTL

Abstract: The thesis focused on the annotation of the chicken genome and the identification of genetic variants to produce GWAS for both macro and expressional phenotypes.
The first objective was to enrich the GRCg7b annotation of the hen genome with lncRNAs, key players in the regulation of gene expression, by combining six multi-tissue projects and reference annotations. The associated GEGA tool enables the expression of these genes to be visualized across 47 tissues and 1,400 samples. Additional work has been carried out on the impact of human assemblies on the detection of aberrant expression involved in rare diseases. Finally, a pipeline to explore the conservation of lncRNAs between species has led to the presumption of orthology links for several hundred of them.
The second objective was to identify SNPs from 700 chicken RNAseq. Their value was assessed for ASE studies, exploring genetic diversity between populations and predicting functional impacts. A pipeline considering the phases of SNPs within a single codon was developed to refine predictions of impact on the protein. At the same time, genotypes obtained at different years and by SNP arrays were analyzed.
Finally, the third objective was to combine genome annotations and identified SNPs to detect regions associated with variability in gene expression (eQTL) and more complex phenotypes (QTL) via the ChickenGTEx project and a population of commercial laying hens.