HAL
open science

Analyse intégrative des ARN longs non-codants chez le chien et leurs implications dans le mélanome oral canin, modèle des mélanomes humains

Céline Le Béguec

To cite this version:

Céline Le Béguec. Analyse intégrative des ARN longs non-codants chez le chien et leurs implications dans le mélanome oral canin, modèle des mélanomes humains. Médecine humaine et pathologie. Université de Rennes, 2018. Français. NNT : 2018REN1B030 . tel-02173770

HAL Id: tel-02173770

https://theses.hal.science/tel-02173770

Submitted on 4 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : « *Génétique, Génomique, Bioinformatique* »

Par
Céline LE BÉGUEC

Analyse intégrative des ARN longs non-codants chez le chien et leurs implications dans le mélanome oral canin, modèle des mélanomes humains

Thèse présentée et soutenue à Rennes, le 10 octobre 2018
Unité de recherche : UMR6290, CNRS-Université Rennes 1

Rapporteurs avant soutenance :

Camille BERTHELOT
Chargée de Recherche,
CNRS UMR8197, INSERM U1024, IBENS

Clément CARRÉ
Maitre de Conférence, UPMC UMR 7622

Composition du Jury :

Camille BERTHELOT — Rapporteure
Chargée de Recherche,
CNRS UMR8197, INSERM U1024, IBENS

Clément CARRÉ — Rapporteur
Maitre de Conférence, UPMC UMR 7622

Marie-Dominique GALIBERT — Examinatrice
Professeure d'universitaire - Praticienne hospitalière,
UMR6290, IGDR, CHU/UR1

Sandrine LAGARRIGUE — Examinatrice
Professeure, Agrocampus Ouest, UMR INRA-AO PEGASE

Denis TAGU — Examinateur
Directeur de Recherche, UMR1349 IGEPP, INRA

Laurent TIRET — Examinateur
Professeur DVM, Université Paris-Est, UMR955 INRA-EnVA

Christophe HITTE — Directeur de thèse
Ingénieur de Recherche, UMR6290, IGDR, CNRS/UR1

Thomas DERRIEN — Co-directeur de thèse
Chargé de Recherche, UMR6290, IGDR, CNRS/UR1

Remerciements

Je tiens à remercier premièrement Camille Berthelot et Clément Carré pour avoir lu attentivement ma thèse, rédigé leurs rapports sur mon travail et autorisé ma soutenance, ainsi que Marie-Dominique Galibert, Sandrine Lagarrigue, Denis Tagu et Laurent Tiret pour avoir accepté d'examiner mon travail.

Je tiens ensuite à remercier grandement Christophe Hitte et Thomas Derrien, mes deux supers directeurs de thèse. Mille mercis pour l'apprentissage, la patience et l'énorme aide tout au long de ces trois années. Christophe:Thomas, un binôme d'encadrement et de co-direction bien huilé. En phase et complémentaire à la fois, ce fut un plaisir de travailler avec vous. L'un comme l'autre amateur de pâtisseries et les premiers à apprécier d'être, comme la majorité de l'équipe, mes béta-testeurs. Je vous remercie tous, j'ai apprécié ces moments de partages.

Je remercie également Catherine André de m'avoir accueillie dans son équipe. Tout au long de ces trois années, j'ai pu rencontrer des collègues/amis au sein de l'équipe. Je pense à Valentin, Stéphanie M., Édouard, Benoît, Aline, Pascale, Frédérique HB., Nadine, Laetitia, Stéphanie R., Florent, Anne-Sophie, Clotilde, Annabelle, Frédérique A., Élodie, Naoual et enfin Maud, la 'grande sœur' rassurante, qui nous aide et nous rassure sur les passages de la thèse plus ou moins stressants. Elles auront fait un passage un peu plus court au sein de l'équipe mais j'ai apprécié leurs compagnies, merci à Sophie et Aurore, je leur tire mon chapeau pour cette reprise d'étude. Même s'ils ne sont pas de l'équipe, ils ont fait partie de mon quotidien, je tiens à remercier aussi Géraldine et les Stéphane(s) D.

Au cours des différentes générations de doctorants dans l'équipe, j'ai pu rencontrer différentes personnalités, de belles rencontres. Tout d'abord, il y avait Morgane, toujours la pêche, le sourire et une énergie débordante. Très complice avec Mélanie, dit Princesse, grande fan de la nature et de ses animaux. Dans un autre style, j'ai fait connaissance avec Ronan, blagues à tous épreuves. Arrivée un peu avant moi pour son M2, j'aurais tout de même été, tout au long de ces trois ans, son binôme de thèse, côte à côte pour le rendu du manuscrit et l'une 5 jours après l'autre pour devenir docteures, certains nous appelaient 'Soline'. Je veux bien sûr parler de Solenne, malgré nos caractères opposés, nous avons pu nous entraider. Je la remercie pour son soutien et lui souhaite une belle réussite pour

l’avenir. Luis aura fait un passage rapide dans l'équipe mais malgré son côté réservé, j'ai pu rencontrer un Costaricain simple et humble. Et pour finir, Anaïs, notre super woman qui fait tout de front, sa thèse, un DU, des remplas et de la cancéro en clinique spécialisée à Paris, admirative, je lui souhaite tout plein de bonne chose pour la suite. Au sein de l’institut, d’autres doctorants auront marqué ces trois ans pour moi, je pense à Hélène, merci pour tous les moments passés en ta compagnie (sorties bijoux, soirées cinéma et weekends entre amis), Karen, Marion D., Marie, Florian et Guillaume.

Petite dédicace particulière à notre chargée de communication, Marion L., bien plus qu’une collègue, elle restera une très bonne amie.

Merci à la Gestion, petite équipe très agréable, Nadine, Isabelle, Rolande, Nathalie et Erwan, sans eux les missions-déplacements-factures en tout genre deviendraient un casse-tête sans nom.

J’espère n’avoir oublié personne dans cette liste.

Je remercie également Frédéric Chalmel, mon tuteur de thèse, Sarah Djebali et Mathieu Emily, membres de mon comité de thèse pour les deux comités réalisés qui m’ont permis de faire un grand bilan, d’avoir un recul sur mon travail, des conseils et des encouragements pour poursuivre mon projet de thèse.

D’un point de vu personnel, je remercie toute ma famille, grand soutien à distance mais très présent quand même. Et pour finir un énorme merci à mon chéri, qui m'a soutenu dans tous mes projets, de l'IUT à aujourd'hui. Pour les quelques mois de rédaction et de pré-soutenance, j'ai eu une aide précieuse au quotidien avec les bons petits plats et les tâches ménagères. Merci encore.

Plan

Résumé .....

Abstract .....

Liste des abréviations .....

Tables des illustrations .....

Introduction **1**

I. Les ARN longs non-codants : lncRNAs *2*

1. Séquençage par RNA-seq 3

1.1. La technologie RNA-seq 3

1.2. L'analyse bioinformatique du RNA-seq 4

2. Caractéristiques des lncRNAs 7

3. Annotation des lncRNAs par RNA-seq 8

4. Rôles des lncRNAs 9

5. Implications des lncRNAs dans le développement des maladies et des cancers 12

6. Conservation des lncRNAs entre espèces 14

7. Les éléments transposables et les lncRNAs 16

7.1. Les principales classes d'éléments transposables 17

7.2. Un élément transposable canin spécifique 19

II. Le chien comme modèle d'étude *19*

1. Les origines, du loup aux chiens modernes 19

2. L'annotation du génome canin 22

3. Les prédispositions du chien à différentes maladies génétiques 24

III. Le chien, modèle spontané des cancers de l'homme, exemple du mélanome muqueux *27*

1. L'intérêt d'un modèle spontané 27

2. L'oncologie comparée entre le chien et l'homme 28

3. Les mélanomes 30

3.1. Les principaux types de mélanomes 30

3.2. Le mélanome muqueux 32

3.2.a. Les mélanocytes des muqueuses 32

3.2.b. Le mélanome muqueux chez l'homme 33

3.2.c. Le mélanome muqueux chez le chien 34

Résultats **38**
I. Identification du répertoire de lncRNAs *38*
1. FEELnc, un programme d'annotation des lncRNAs 38
2. FEELnc et son application au transcriptome canin 40
3. Intersection de l'annotation Canfam3.1-plus et des promoteurs et enhancers 41
II. Caractérisation des lncRNAs canins *42*
1. Caractérisation des profils d'expression des lncRNA canins 42
1.1. Analyse comparée des niveaux d'expression des lncRNAs et des mRNAs 43
1.2. Analyse de la tissu-spécificité des lncRNAs 44
1.3. Analyse des données d'expression par clustering ascendant hiérarchique 45
1.4. Corrélation entre la tissu-spécificité des lncRNAs et leur contenu en Éléments Transposables (TEs) 47
2. Prédiction fonctionnelle bioinformatique des lncRNAs canins 48
2.1. Analyse des relations d'orthologie des lncRNAs entre l'homme et le chien 48
2.1.a. La génomique comparative 48
2.1.b. Apport de la transcriptomique comparative 50
2.2. Analyse de co-expression entre lncRNAs et mRNAs canins 52
2.2.a. Distance et type de la relation de co-expression 53
2.2.b. Caractérisation par ontologies de gènes des lncRNAs impliqués dans les co-expressions. 54
III. Implication des lncRNAs dans le mélanome oral canin *54*
1. Contrôle qualité des échantillons de mélanomes oraux 55
2. Les lncRNAs différentiellement exprimés entre tumeurs et contrôles 56
2.1. DESeq2 versus EBSeq 56
2.2. DESeq2 versus $DESeq2_{multifacteur}$ 57
3. Des lncRNAs biomarqueurs potentiels 58
Discussion - Perspectives **60**
1. L'optimisation du dispositif expérimental 60
2. Génomique comparative et conservation des lncRNAs 62
3. Approches complémentaires au RNA-seq 63
4. La co-expression lncRNA:mRNA 64
5. Le mélanome oral canin 66
6. La validation fonctionnelle 68
Conclusion **69**

Matériels et Méthodes **71**
I. Identification et caractérisation des lncRNAs canins *71*
1. Données initiales issues de 26 RNA-seq 71
2. Description de l'annotation de référence du génome canin (CanFam3.1-plus) 72
3. Intersection de l'annotation Canfam3.1-plus et des promoteurs et enhancers 72
4. Pipeline d'analyse bioinformatique des données de RNA-seq 73
5. Calcul de la tissu-spécificité d'un gène 74
6. Clustering ascendant hiérarchique 75
7. Analyse du contenu en Éléments Transposables (TEs) 75
8. Le contenu en TE et l'expression des gènes 75
9. Génomique comparative et conservation évolutive des lncRNAs / mRNAs canins 76
9.1. Approche par synténie 76
9.2. Par utilisation du score GERP 76
10. Transcriptomique comparative et conservation des profils d'expression chien / homme 77
11. Représentation circulaire des données de transcriptomique comparative avec Circos 77
12. Analyse de la co-expression des paires lncRNA:mRNAs canins 78
13. Analyse des enrichissement en termes GO ('Gene Ontology') 79
II. Implication des lncRNAs dans le mélanome oral canin *79*
1. Données initiales issues de 60 RNA-seq 79
2. Recherche de lncRNAs différentiellement exprimés 80
2.1. L'outil DESeq2 80
2.2. L'outil EBSeq 81
Bibliographie **82**
Annexes **119**
I. Annexe 1 : Characterization and functional predictions of canine long non-coding RNAs, Le Béguec et al., 2018 publié dans Scientific Reports. *119*
II. Annexe 2 : FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome, Wucher et al., 2017 publié dans Nucleic Acids Research. *119*

Résumé

Les ARN longs non-codants (lncRNAs) constituent une famille d'ARN hétérogènes qui jouent un rôle majeur dans de nombreux cancers et notamment dans les mélanomes. Le chien est un modèle naturel et spontané pour l'analyse génétique comparée des cancers et, l'annotation du génome canin a récemment été enrichie avec l'identification de plus 10 000 lncRNAs. Afin de réaliser des prédictions fonctionnelles bioinformatiques des lncRNAs, nous avons caractérisé les profils d'expression des lncRNAs canins à partir de 26 tissus distincts. Nous avons défini la spécificité tissulaire de l'expression des lncRNAs et inféré leur fonctionnalité potentielle par des analyses de génomique et de transcriptomique comparatives avec des données humaines issues du projet ENCODE (ENCyclopedia Of DNA Elements). Comme chez l'homme et la souris, une grande proportion de lncRNAs canins (44 %) est exprimée de manière spécifique au sein d'un tissu. Par une approche de génomique comparative, nous avons identifié plus de 900 lncRNAs orthologues entre l'homme et le chien et pour 26 % d'entre eux, des patrons d'expression entre tissus significativement conservés ($p < 0,05$).

Dans le cadre de l'étude des mélanomes canins, nous avons analysé les données de RNA-seq de 52 échantillons tumeurs/contrôles de mélanomes oraux. Nous avons identifié plus de 750 lncRNAs différentiellement exprimés entre la tumeur et le contrôle (FDR < 0,01), dont plus de 100 conservés avec l'homme. Ces lncRNAs constituent de bons candidats pour étudier la régulation de la progression tumorale des mélanomes chez le chien et pourront être évalués pour leurs potentiels diagnostic et thérapeutique en médecine humaine et vétérinaire.

Abstract

Long non-coding RNAs (lncRNAs) are a family of heterogeneous RNAs that play a major role in many cancers, particularly in melanomas. The dog is a natural and spontaneous model for the comparative genetic analysis of cancers and, the annotation of the canine genome has recently been enriched with the identification of over 10,000 lncRNAs. In order to perform functional bioinformatic predictions of lncRNAs, we have characterized the expression patterns of canine lncRNAs from 26 distinct tissues representative of the major functions of the organism. We defined the tissue specificity of lncRNAs expression and inferred their potential functionality by comparative genomic and transcriptomic analyses with human data from the ENCODE project (ENCyclopedia Of DNA Elements). As in humans and mice, we show that a large proportion of canine lncRNAs (44%) are expressed specifically within a tissue. Using a comparative genomic approach, we have identified more than 900 orthologue lncRNAs between humans and dogs, and we show that for 26% of them, tissue expression patterns are also significantly conserved ($p < 0.05$).

In the study of canine melanomas, we investigated the lncRNAs from RNA-seq data from 52 tumour/control samples of oral melanoma. We identified more than 750 lncRNAs differentially expressed between tumour and control (FDR < 0.01), of which more than 100 were conserved with humans. These lncRNAs are good candidates to study the regulation of tumour progression of melanomas in dogs and can be evaluated for their diagnostic and therapeutic potential in human and veterinary medicine.

Liste des abréviations

ACP	Analyse en Composante Principale
API	Application Programming Interface
BAM	'Binary Alignment Map'
CAGE	'Cap Analysis of Gene Expression'
ChIP-Seq	Immunoprécipitation de la chromatine-Séquençage
CPS	'Coding Potential Score'
CRB Cani-DNA	Centre de Ressources Biologiques Cani-DNA
CTC	'Circulating Tumours Cells'
DL	Déséquilibre de Liaison
EGFR	*'Epidermal Growth Factor Reception'*
EPO	Enredo-Pecan-Ortheus
EST	Étiquettes de Séquences Transcrites
FDR	'False Discovery Rate'
FPKM	Fragments Par Kilobase d'exon par Million de fragments mappés
GO	'Gene Ontology'
GRET	Golden Retriever
H3K27ac	Marque d'acétylation de la lysine 27 de l'histone 3
HSAN	Neuropathies héréditaires sensitives et autonomes
Kb	Kilobase
lFC	'Log Folde Change'
LINE	'Long INterspersed Elements'
LOF	Livre des Origines Français
LTR-ERVL	'Long Terminal Repeat-Endogenous RetroVirus-Like elements'
MALAT1	*'Metastasis-associated lung adenocarcinoma transcript 1'*
Mb	Mégabase

ADN	Acide DésoxyriboNucléique
ARN	Acide RiboNucléique
BH	Benjamini-Hochberg
CASC15	*'Cancer Susceptibility Candidate gene 15'*
CNV	'Copy Number Variation'
CPU	'Central Processing Unit'
CS	'Conserved Segment'
DE	Différentiellement Exprimés
EBI	'European Bioinformatic Institute'
ENCODE	'ENcyclopedia of Dna Elements'
eRNA	enhancer RNA
FCI	Fédération Cynologique Internationale
FEELnc	'Flexible Extraction of LncRNAs'
GERP	'Genomic Evolutionary Rate Profiling'
GOLD	'Genomes OnLine Database'
GTEx	'Genotype-Tissue Expression'
H3K4me3	Marques de tri-méthylation de la lysine 4 de l'histone 3
HULC	*'Highly up-regulated in liver cancer'*
LABR	Labrador
lincRNA	long ARN non-codants intergéniques
lncRNA	long ARN non-codants
LTR	'Long Terminal Repeats'
MA	Millions d'Années
MAPK	'Mitogen activated protein kinase'
MEG3	*'Maternally expressed 3'*

mRNA	Gènes codant pour des protéines
NGS	'Next Generation Sequencing'
NMD	'Non-Sense Mediated Decay'
NRON	*'Non-coding Repressor Of NFAT'*
OMIA	'Online Mendelian Inheritance in Animals'
pancRNA	'Promoter Associated ncRNAs'
PCA3	'Prostate Cancer Antigen'
POPDC3	'Popeye Domain Containing 3'
RH	Cartes Hybride d'Irradiation
ROC	'Receiver Operating Characteristic'
SAM	'Sequence Alignment Map'
SINE	'Short Interspersed Elements'
Sn	Sensibilité
Sp	Spécificité
TE	Elément transposable
TLS	*'Transfer RNA-like structures'*
TPM	Transcrits Par Million
TSS	Sites d'initiation de la transcription
UCLA	Université de Californie à Los Angeles
WGA	'Whole Genome Alignment'
Xist	*'X Inactive Specific Transcript'*

NFAT	Facteur nucléaire des cellules T activées
NIH	'National Institutes of Health'
NRG3	*Neuregulin 3*
nt	Nucléotide
ORF	Cadres Ouverts de Lecture
pb	Paire de base
PODL	'Poodle' (Caniche)
RF	Random Forest
RNA-seq	RiboNucleic Acid Sequencing
RPKM	reads par kilobase d'exon par
SCC	Société Centrale Canine
SINEC_Cf	SINE Carnivore_Canis familiaris
SNP	'Single-Nucleotide Polymorphism'
SSM	'Superficial Spreading Melanoma'
TFBS	Site de liaison aux facteurs de transcriptions
TP53	*'Tumor Protein 53'*
TRDN	*Triadine*
TUCp	'Transcript of Unknown Coding potential'
UV	Ultra-Violets
WGCNA	Analyse de réseau de co-expression génique pondérée
ZEB2	*'Zinc Finger E-Box homeobox 2'*

Table des illustrations

Figure 1 : Pipeline bioinformatique de RNA-seq 6

Figure 2 : Fonctions des longs ARN non-codants 10

Figure 3 : Insertion du TE carbonaria dans un intron du gène du cortex chez le phalène du bouleau. 17

Figure 4 : Schéma des principales classes d'éléments transposables (TEs) et de leur organisation génétique typique 17

Figure 5 : Représentation schématique des deux goulets d'étranglement de l'histoire du chien 20

Figure 6 : Phylogénie des races de chiens 21

Figure 7 : Avancée de la génétique humaine et canine 22

Figure 8 : Arbre phylogénétique 22

Figure 9 : La diversité des races de chiens 24

Figure 10 : Schéma de la muqueuse 33

Figure 11 : Description de FEELnc$_{_codpot}$ et FEELnc$_{_classifier}$ 40

Figure 12 : Analyse comparative des niveaux de transcription entre les mRNAs et les lncRNAs 43

Figure 13 : Analyse comparative des niveaux de transcription entre les antisense et les lincRNAs 43

Figure 14 : Proportion des lncRNAs tissus-spécifique et des mRNAs tissus-spécifique 44

Figure 15 : Clustering hiérarchiques ascendants des 26 tissus 45

Figure 16 : Distribution du niveau d'expression de NRG3 et son antisense 46

Figure 17 : Répartition et enrichissement des TEs dans le génome 47

Figure 18 : Corrélations canoniques entre l'expression des gènes dans les 26 tissus 48

Figure 19 : Diagramme Circos fournissant la visualisation des relations orthologues humaines des lncRNAs et de leurs modèles d'expression transcriptomique comparée 49

Figure 20 : Diagrammes à barres représentant l'expression des 11 tissus pour le chien et l'homme.. 51

Figure 21 : Corrélation de la paire lncRNA:EGFR 53

Figure 22 : Schémas de co-expression des paires la paire lncRNA:mRNA 53

Figure 23 : Analyse en composante principale des 60 échantillons initiaux de tumeurs de mélanome oral canin 55

Figure 24 : Analyse des niveaux de transcription des lncRNAs dans les 52 échantillons de tumeurs de mélanome oral canin 56

Figure 25 : Diagramme de Venn permettant de visualiser le nombre de lncRNAs différentiellement exprimés (DE) en commun entre ceux issus de DESeq2 et ceux issus d'EBSeq 56

Figure 26 : Diagramme de Venn permettant de visualiser le nombre de lncRNAs différentiellement exprimés (DE) en commun entre ceux issus de DESeq2$_{_défaut}$ et ceux issus de DESeq2$_{_multifacteur}$ 58

Figure 27 : Deux exemples de gènes DE 58

Figure 28 : Schémas représentant les antisenses (à gauche) et les lincRNAs (à droite) 78

Table 1 : Comparaison des annotations CanFam3.1 versus CanFam3.1-plus 23
Table 2 : Exemples de cancers associés à des races de chien spécifiques 27
Table 3 : Description de chaque type de tissu 42
Table 4 : TPM_{moyen} pour chaque tissu pour les lncRNAs et les mRNAs 43
Table 5 : Nombre de lncRNAs exprimés à > 1 TPM et > 0,1 TPM chez le chien et l'homme 50
Table 6 : Analyse d'enrichissement de termes GO 54
Table 7 : Identifiants de lncRNAs chien – humain et noms de gènes humain des 76 lncRNAs différentiellement exprimés et conservés entre le chien et l'homme 59

Introduction

La génétique est l'étude des gènes, des variations génétiques et de l'hérédité des organismes vivants. Elle est généralement considérée comme un domaine de la biologie, mais recoupe souvent de nombreuses autres sciences de la vie, ainsi que l'informatique et les statistiques. L'origine de la génétique vient du scientifique Gregor Mendel, à la fin du XIXème siècle. En étudiant 'l'héritage des traits' et leurs schémas de transmission des parents à la descendance, il a observé que les organismes (plantes de pois dans son expérience) héritent des traits des parents par le biais de discrètes 'unités d'héritage'. Ce terme, encore utilisé aujourd'hui, est une définition quelque peu ambiguë de ce que nous appelons un gène. Près d'un siècle plus tard, en 1953, la découverte de la structure en double hélice de l'ADN (Watson and Crick, 1953) où les quatre bases (Adénine, Thymine, Guanine et Cytosine) étaient liées par des liaisons hydrogènes, a permis le développement de plusieurs techniques de séquençage d'ADN, qui se succèdent au cours des années. En particulier en 1977, Frederick Sanger a mis au point une méthode de séquençage nommée par son nom, la méthode de Sanger (Sanger, Nicklen and Coulson, 1977). Cette méthode, toujours employée aujourd'hui, a permis la réalisation de la séquence du premier génome complet d'un organisme biologique, le virus bactériophage φX174 (Sanger, Nicklen and Coulson, 1977). Un peu plus tard, dans les années 90, ce sont les puces à ADN qui apparaissent. Elles permettent notamment d'analyser le niveau d'expression des gènes, mais restent limitées par les sondes présentes. C'est en 2001 que la séquence complète du génome de l'homme est publiée (Lander *et al.*, 2001; Venter *et al.*, 2001), suivie de celles de la souris, du chimpanzé, du rat et en 2005, de celle du génome du chien (Lindblad-Toh *et al.*, 2005).

Depuis une dizaine d'années, les technologies de séquençage dites de nouvelles générations (NGS, 'Next Generation Sequencing') ou encore haut-débit ont atteint un débit et une précision tels qu'elles permettent aux chercheurs de mettre en œuvre des projets de génomique de très grande envergure en un temps raisonnable et des coûts abordables. Les approches NGS permettent une parallélisation de plusieurs échantillons sur une même expérimentation, une meilleure sensibilité et un séquençage plus rapide (Ilie *et al.*, 2014). Toutes ces technologies ont permis de générer de plus en plus de données pour différents organismes.

En revanche, ces technologies et l'accumulation de projets visant à réaliser entre autre un génome complet, mais également de produire des données épigénétiques avec le ChIP-seq, de transcriptomique avec le RNA-seq ou encore d'exomes, engendrent une grande quantité de données qu'il faut stocker, analyser et mettre à disposition de la communauté scientifique. Cette étape de mise à disposition des données est assurée par les bases de données publiques telles que GenBank (Benson *et al.*, 2013), Ensembl (Yates *et al.*, 2016), UCSC (Casper *et al.*, 2018), SRA (Alnasir and Shanahan, 2015), etc. En 2018, la base de données GOLD ('Genomes OnLine Database'[1]), dénombre plus de 200 000 projets de séquençage (complets ou partiels) et près de 300 000 organismes sont recensés ; appartenant aux archées, bactéries, procaryotes et eucaryotes, ils disposent d'un séquençage complet de leurs génomes. Ce chiffre augmentant chaque année, la question du stockage est désormais essentielle pour assurer la sécurité, la disponibilité et la diffusion des données.

I. Les ARN longs non-codants : lncRNAs

Entre 60 et 70 % du génome humain est transcrit mais seulement 1 à 2 % de ces ARN, les ARN messagers (mRNAs) codent pour des protéines (ENCODE Project Consortium, 2012). Les ARN messagers sont étudiés depuis de nombreuses années et leurs caractéristiques ainsi que leurs rôles sont définis pour une grande partie même s'il reste encore de nombreuses questions à investiguer concernant leurs structures et leurs régulations. De façon complémentaire aux mRNAs, la proportion importante d'ARN transcrits mais non codant pour des protéines a longtemps été associée à de l'ADN 'poubelle' voire à du bruit de fond transcriptionnel. Cependant, depuis la découverte des petits ARN non-codants (ARN de transfert, microRNA…) jusqu'aux études plus récentes réalisées dans le cadre du consortium ENCODE ('Encyclopedia of DNA Elements') (Dunham et al., 2012), nous considérons actuellement que l'annotation du génome humain est composée d'environ 60 000 gènes de différentes classes (ou biotypes), correspondant à plus de 20 000 gènes codant pour des protéines, ~ 16 000 ARN longs non-codants, ~ 10 000 petits ARN non-codants et

[1] https://gold.jgi.doe.gov/

~ 14 000 pseudogènes (GENCODE version 28[2]) (Harrow *et al.*, 2012). Le biotype, défini ici comme une classe d'ARN (mRNA, lncRNA, miRNA, rRNA…), sur lequel nous avons concentré nos analyses dans ce projet de thèse est celui des ARN longs non-codants (lncRNAs), représentant une portion importante du génome de mammifère, mais encore trop peu caractérisée (Signal, Gloss and Dinger, 2016).

1. Séquençage par RNA-seq

1.1. La technologie RNA-seq

Décrypter le transcriptome, c'est-à-dire l'ensemble des molécules d'ARN d'une cellule ou d'un tissu, est essentiel pour interpréter les éléments fonctionnels du génome, ainsi que pour comprendre le développement de l'organisme et les différentes maladies. Pour cela, dès les années 90, des puces à ADN ont permis d'analyser les profils d'expression dans les cellules, tissus, organes ou organismes. Cette technologie utilise la propriété de l'ADN dénaturé qui consiste à reformer spontanément sa double hélice lorsqu'il est en présence d'un brin complémentaire (Schena *et al.*, 1995). Malgré les nombreuses applications, telles que le diagnostic (Batzir, Shohat and Maya, 2015), le criblage de mutation (Ahrendt *et al.*, 1999) ou l'étude phylogénétique (Cho and Tiedje, 2001), les puces à ADN ont leurs limites notamment de n'interroger que les gènes représentés sur les puces.

C'est en 2005 qu'apparaît la nouvelle génération de séquençage, dit séquençage de haut-débit (Carninci *et al.*, 2005). Parmi ces nouvelles technologies, le séquençage du transcriptome à haut-débit, le RNA-seq, consiste en une première étape de conversion d'une population d'ARN en une librairie de fragments d'ADNc avec des adaptateurs fixés à une ou aux deux extrémités. La réaction de séquençage est alors réalisée directement sur le support où l'ADNc a été amplifié. Un laser détecte les bases incorporées au fur et à mesure. La lecture, que nous appellerons 'read' au cours de ce manuscrit, est construite ainsi cycle après cycle pour obtenir des séquences courtes à partir d'une extrémité de la molécule (séquençage 'single-end') ou des deux extrémités (séquençage 'paired-end'). Les reads ont généralement une taille de 50 à 400 pb, selon la technologie de séquençage de l'ADN utilisée. Les reads résultantes sont soit alignées sur un génome de référence et/ou des

[2] https://www.gencodegenes.org/stats/current.html

transcrits de référence, ou assemblées *de novo* sans séquence génomique de référence pour produire une véritable cartographie de l'ensemble des molécules d'ARN transcrites à l'échelle de l'échantillon considéré. L'analyse des données va permettre de caractériser à la fois la structure des transcrits et le niveau d'expression pour chaque gène (Wang, Gerstein and Snyder, 2009; Ilie *et al.*, 2014).

Cette technologie possède de multiples avantages, d'une part, des niveaux élevés de reproductibilité, à la fois pour les répliquas techniques et biologiques. D'autre part, l'absence d'étape de clonage nécessitant une quantité moindre d'échantillon d'ARN entraine une baisse des coûts, une diminution des temps de séquençage et une plus grande sensibilité (Wang, Gerstein and Snyder, 2009; Ilie *et al.*, 2014). En revanche, le RNA-seq implique de nombreux défis informatiques, au niveau matériel, il nécessite des ordinateurs ou des serveurs avec une grande capacité de calculs et de stockage et également au niveau du traitement informatique et statistique et à l'interprétation des données pour limiter les erreurs (Behjati and Tarpey, 2013).

1.2. L'analyse bioinformatique du RNA-seq

Le séquençage de l'ARN (RNA-seq) est devenu l'une des technologies les plus largement utilisées dans la recherche biomédicale pour des mesures d'expression des transcrits. À partir du séquençage, les données doivent être traitées de manière très spécifique. À cette étape, les données sont sous forme de fichier .fastq, permettant le stockage des séquences ainsi que leurs scores de qualité. De nombreuses suites logicielles existent pour réaliser les principales étapes de l'analyse bioinformatique des données RNA-seq : l'alignement (ou 'mapping'), la reconstruction des modèles de gènes, la quantification des transcrits/gènes, voire l'analyse d'expression différentielle.

Dans un premier temps, l'étape de mapping de séquences ARN peut être compliquée par des facteurs biologiques (polymorphisme, épissage alternatif/pathologique) et/ou techniques (erreurs de séquençage, séquences de faible complexité, une annotation incomplète, etc.). Pour qu'un aligneur RNA-seq soit efficace, il doit (i) aligner les reads en prenant en compte les jonctions d'épissage annotées ou non, (ii) gérer les reads appariées, (iii) gérer les données spécifiques au brin (protocole de construction de banque spécifique de brin) et (iv) fonctionner efficacement. Rien que pour l'étape de mapping, plus d'une

quinzaine de programmes ont été créés et certains d'entre eux ont fait l'objet d'études comparatives. Parmi le large panel de logiciels testés dans une étude comparative de Baruzzo *et al.* de 2017, trois d'entre eux (CLC[3], STAR (Dobin *et al.*, 2013) et Novoalign[4]) semblent plus précis lors de l'alignement de read unique et identifient correctement les limites d'intron.

Dans un second temps, il s'agit de réaliser une reconstruction des modèles de transcrits en se basant sur les cumuls des alignements de reads sur le génome de référence étudié. En revanche, certains programmes adoptent l'approche alternative de la reconstruction *de novo*, où des séquences de transcription contiguës sont assemblées sans utiliser de génome de référence (Haas *et al.*, 2013). Pour améliorer la précision de l'annotation existante en utilisant des données RNA-seq, Cufflinks (Roberts *et al.*, 2011) et iReckon (Mezlini *et al.*, 2013) par exemple, consultent les structures de gènes connues pendant le processus d'assemblage des transcrits et peuvent être utiles pour affiner les coordonnées des jonctions d'exons et de début et fin de transcription.

Dans un troisième temps, la quantification des transcrits/gènes est réalisée. De la même façon que pour les aligneurs, il existe une multitude d'outils bioinformatiques. Un des premiers défis rencontrés par l'étude de Teng *et al.*, de 2016 (Teng *et al.*, 2016), réalisée sur la comparaison des performances entre les différentes approches, est l'absence d'unité standard pour la quantification du niveau de transcription. Par exemple, Cufflinks (Trapnell *et al.*, 2012) rapporte des fragments par kilobase d'exon par million de fragments mappés (FPKM) ; Flux Capacitor (Montgomery *et al.*, 2010) rapporte des reads par kilobase d'exon par millions de reads lues (RPKM) ; eXpress (Roberts and Pachter, 2013), RSEM (Li and Dewey, 2011), Sailfish (Patro, Mount and Kingsford, 2014), kallisto (Bray *et al.*, 2016), et Salmon (Patro, Duggal and Kingsford, 2015) rapportent des transcrits par million (TPM). En effet, une normalisation efficace est nécessaire pour comparer les profils d'expression inter-échantillons afin de prendre en compte la profondeur de séquençage et la taille des transcrits de l'annotation de référence. Un critère important lors d'un traitement de données, est la vitesse d'exécution et en ce sens, l'outil IsoEM (Nicolae *et al.*, 2011) semble

[3] https://www.qiagenbioinformatics.com/products/clc-genomics-workbench
[4] http://www.novocraft.com/products/novoalign/

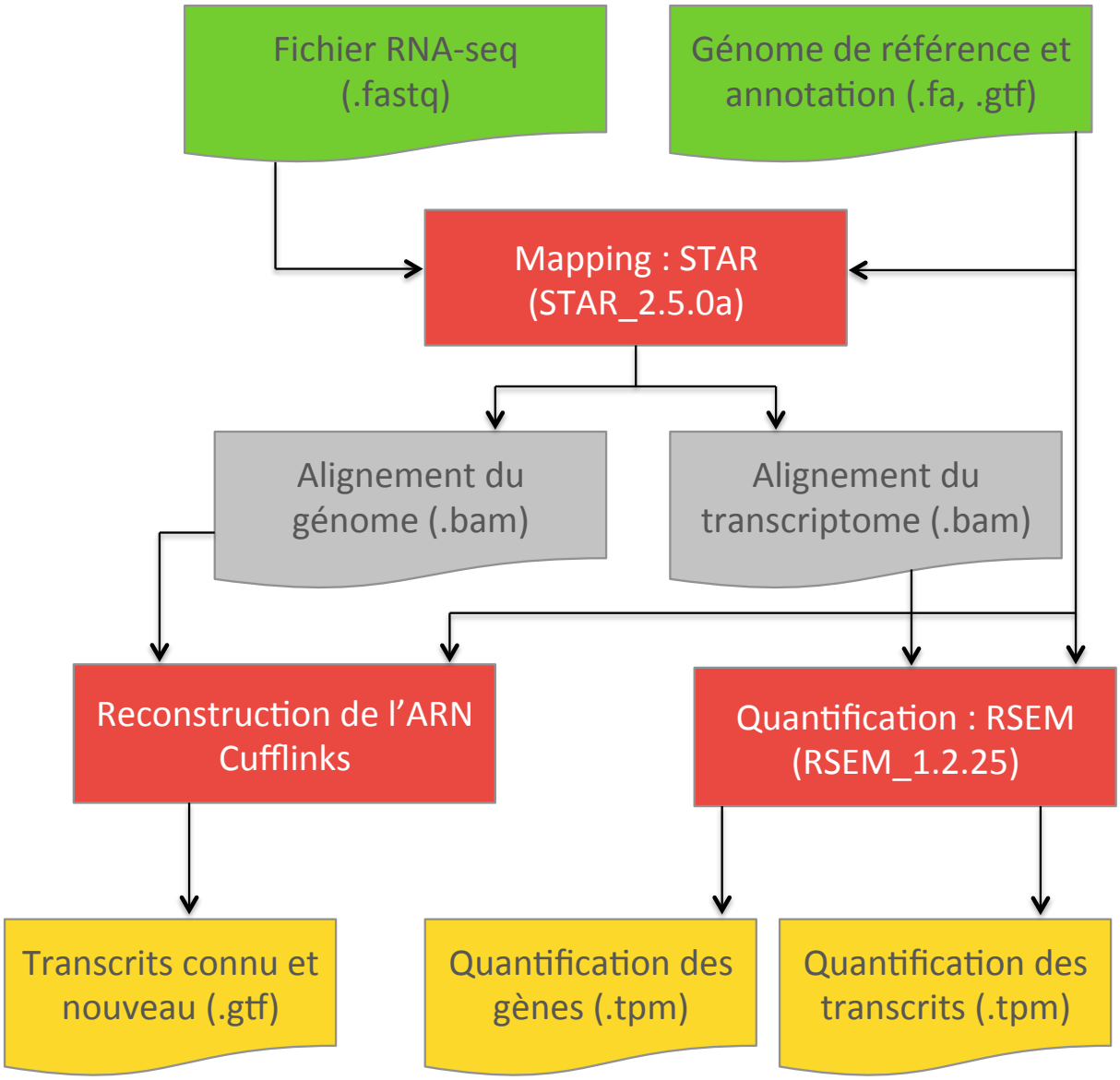

Figure 1 : Pipeline bioinformatique de RNA-seq*, schéma traduit de Djebali et al., 2017*

le plus rapide dans l'étude de Chandramohan *et al*. de 2013 (Chandramohan *et al.*, 2013). Il parvient à tirer le meilleur parti des ressources disponibles en occupant tous les CPU ('Central Processing Unit') disponibles. Pour Cufflinks et RSEM, il est également possible d'accélérer le processus de quantification grâce au réglage 'multi-thread', qui consiste à exécuter plusieurs tâches en parallèle. En revanche des outils comme HTSeq[5], ne supportent pas la parallélisation des calculs. Enfin, Cufflinks, RSEM et IsoEM présentent des valeurs d'expression avec une grande précision.

Les données de RNA-seq permettent également de réaliser des analyses d'expression différentielle (DE) qui dans le cas de maladies, permettent d'effectuer la comparaison de l'expression différentielle d'un ensemble de gènes entre des cas et des contrôles. Différentes méthodes d'analyse de différentiel d'expression sont mises à disposition lors de ces études. Nous retrouvons des méthodes qui modélisent l'expression des gènes selon la loi binomiale négative tels que edgeR (Robinson, McCarthy and Smyth, 2010) ou DESeq (Love, Huber and Anders, 2014), des approches non paramétriques telles que NOIseq (Tarazona *et al.*, 2011) et SAMseq (Li and Tibshirani, 2013), des transformations de comptages de read au niveau des gènes pour la modélisation linéaire avec limma (Smyth, 2004), ainsi que des méthodes de détection basées sur les isoformes, comme Cuffdiff 2 (Anders and Huber, 2010) et EBSeq (Leng *et al.*, 2013).

Généralement, l'ensemble des étapes bioinformatiques est réuni dans un pipeline (Figure 1) (Djebali *et al.*, 2017), tel que celui inspiré du pipeline ENCODE[6], qui assure les analyses d'alignements et de reconstruction des transcrits à partir de données RNA-seq. Dans ce pipeline, l'étape de mapping est réalisée à l'aide de l'outil STAR (voir Matériels et Méthodes) et consiste à aligner les séquences issues du séquençage avec le génome de référence. Cette étape permet de générer des fichiers .bam ('Binary Alignment Map'), correspondant à la compression en format binaire des fichiers .sam ('Sequence Alignment Map'), supportant les séquences alignées (Djebali *et al.*, 2017). Parmi les outils présentant de très bonnes caractéristiques pour la reconstruction ainsi que la quantification, Cufflinks et RSEM sont utilisés dans ce pipeline (Figure 1). Ces étapes permettent d'obtenir des transcrits

[5] http://htseq.readthedocs.io/en/release_0.10.0/
[6] https://github.com/ENCODE-DCC/long-rna-seq-pipeline/blob/master/DAC/STAR_RSEM.sh

connus et/ou nouveaux et de déterminer l'expression des transcrits ou gènes d'un échantillon à un temps donné.

2. Caractéristiques des lncRNAs

Un ARN long non-codant est un ARN présentant plus de 200 nucléotides (nt) et contrairement aux gènes codant pour les protéines (mRNAs), n'est pas traduit en protéine dans la grande majorité des cas. À l'instar des mRNAs, les lncRNAs sont souvent transcrits par l'ARN polymérase II, subissent l'ajout d'une coiffe et peuvent être polyadénylés, mais ne possèdent pas ou très rarement la capacité à coder des protéines ou, en d'autres termes, leur 'potentiel codant' est faible. Chez l'homme, il a été montré que la plupart des lncRNAs sont épissés (98 %) et présentent deux exons pour 42 % d'entre eux contre seulement 6 % pour les mRNAs (Derrien, Johnson, *et al.*, 2012). La différence entre ces deux biotypes (lncRNA et mRNA) se retrouve également dans la taille des exons et des introns, qui sont plus grands pour les lncRNAs que pour les mRNAs (Derrien, Johnson, *et al.*, 2012). En revanche, la taille des transcrits des lncRNAs est plus petite que celle des mRNAs (médiane de 592 nt pour les lncRNAs contre 2 453 nt pour les mRNAs), dû au fait qu'ils possèdent moins d'exons (Derrien, Johnson, *et al.*, 2012; Al-Tobasei, Paneru and Salem, 2016). Ces analyses ont récemment été remises en question via le développement d'une stratégie combinant capture d'ARN et séquençage par longues reads de type PacBio (Lagarde *et al.*, 2017). Dans cette étude, les auteurs montrent que la médiane de taille des lncRNAs (~ 1 300 nt) est deux fois plus grande que celle préalablement établie et suggèrent que la différence de taille entre lncRNA et mRNA est simplement liée au protocole de séquençage impliquant une étape de sélection par taille des fragments d'ARN.

Les lncRNAs possèdent un turnover bien plus rapide que celui des mRNAs (Kutter *et al.*, 2012), c'est-à-dire qu'une séquence de lncRNA peut acquérir une nouvelle fonction ou perdre sa fonction plus rapidement que les mRNAs, ce qui a pour conséquence directe que leurs séquences sont peu conservées comparativement aux mRNAs (Soumillon *et al.*, 2013; Kornienko *et al.*, 2016). Ainsi, les nouveaux lncRNAs émergent au cours de l'évolution selon trois principaux mécanismes : (i) formation *de novo* à partir d'une séquence génomique non transcrite précédemment, (ii) duplication d'un autre lncRNA, ou encore (iii) transformation d'un gène codant pour une protéine en lncRNA ou pseudogénisation (Ulitsky *et al.*, 2011).

Généralement, les lncRNAs sont répartis en deux groupes selon leurs positions génomiques et leurs orientations de transcription. Ainsi, nous pouvons distinguer les lncRNAs *antisens*, qui correspondent à des lncRNAs chevauchant un ou plusieurs mRNAs mais transcrit dans le sens opposé de ce(s) dernier(s) et, les longs ARN non-codants intergéniques (*lincRNAs* pour 'long intergenic ncRNAs' ou 'large intervening ncRNAs' (Guttman *et al.*, 2009)) qui sont, localisés entre les positions des mRNAs.

3. Annotation des lncRNAs par RNA-seq

Au début des années 2000, des études génomiques ont montré qu'une grande partie du génome des mammifères peut être transcrite (Carninci *et al.*, 2005), suggérant la présence d'un nombre important de transcrits non-codants, en comparaison des 1 à 2 % de séquences codant pour des protéines. Cette mise en évidence a nécessité d'importantes investigations pour cataloguer les éléments non-codants, à l'aide de données recueillies via des puces à ADN (Bertone *et al.*, 2004; Kapranov *et al.*, 2007), des productions d'étiquettes de séquences transcrites (EST), de l'ADNc cloné (Carninci *et al.*, 2005; ENCODE Project Consortium *et al.*, 2007) ou encore des cartographies des modifications d'histones (Guttman *et al.*, 2009), et depuis quelques années, l'annotation des lncRNAs par RNA-seq (Cabili *et al.*, 2011; Derrien, Johnson, *et al.*, 2012).

Chez l'homme, un premier catalogue de 8 195 lincRNAs putatifs a pu être identifié automatiquement via l'intégration des données RNA-seq de 24 tissus et types de cellules avec des annotations de transcription disponibles à la communauté (Cabili *et al.*, 2011). Près de 58 % de ces transcrits apparaissent comme nouveaux car identifiés pour la première fois en utilisant du RNA-seq (Cabili *et al.*, 2011). Chacun de ces lincRNAs a été défini grâce à un ensemble de caractéristiques structurelles, séquentielles et d'expression et 993 de ces lincRNAs, soit 12 %, présentent des transcrits orthologues exprimés dans d'autres espèces de vertébrés.

Depuis 10 ans, le RNA-seq a permis l'identification de catalogues de lncRNAs dans certaines espèces animales d'intérêts. Nous retrouvons par exemple, dans l'étude de Zhang *et al.*, de 2014 (Zhang *et al.*, 2014) l'identification de lncRNAs dans le développement embryonnaire, où un ensemble de 5 563 transcrits provenant de 3 492 loci ont été identifiés comme nouveaux lncRNAs impliqués dans le développement embryonnaire de la souris.

Une étape importante dans l'annotation des lncRNAs consiste à prédire si les nouveaux modèles de gènes assemblés par RNA-seq suivant l'étape de reconstruction de transcriptome sont codants (mRNAs) ou non-codants (lncRNAs). Afin d'automatiser cette tâche, un outil bioinformatique nommé FEELnc ('Flexible Extraction of LncRNAs') (Wucher *et al.*, 2017) a récemment été développé dans l'équipe, et permet, entre autre, de discriminer les ARN longs non-codants des gènes codant pour des protéines. FEELnc est un programme rapide, car il ne nécessite pas d'alignements ('alignment-free'), et se base sur un modèle d'apprentissage automatique de type forêt aléatoire ('Random Forest') (Breiman, 2001) entrainé avec des caractéristiques inhérentes aux séquences codantes et non-codantes telles que les fréquences de multi-*k*-mer et les cadres de lecture ouverts (ORF) relâchés. FEELnc a été utilisé dans différentes études, notamment la ré-annotation du génome canin (Wucher *et al.*, 2017), présentant un catalogue de 10 444 lncRNAs. Une étude sur le poulet (Muret *et al.*, 2017) a également pu identifier un set de 2 193 genes lncRNAs putatifs dont 1 670 étaient exprimés de façon reproductible dans les tissus adipeux et/ou foie. Enfin, FEELnc a aussi été utilisé sur des données RNA-seq de l'algue brune (*Ectocarpus*) aboutissant à l'un des premiers catalogues de lncRNAs identifiés chez l'algue comprenant 717 lncRNAs (Cormier *et al.*, 2017).

4. Rôles des lncRNAs

Les lncRNAs ont longtemps été associés à du bruit de fond de transcription, mais depuis quelques années, avec l'efficacité, la résolution du séquençage haut-débit et au nombre croissant de tissus utilisés, les analyses de lncRNAs ont permis d'en annoter un plus grand nombre et de mettre en évidence leurs implications dans un grand nombre de processus biologiques, notamment dans beaucoup de cancers (Leucci, Vendramin, *et al.*, 2016). La découverte et l'étude des lncRNAs sont d'une importance majeure pour la biologie et la compréhension des maladies puisqu'ils représentent une composante extensive du génome qui reste encore largement inexplorée (Mattick, 2009; Ponting, Oliver and Reik, 2009).

Ainsi, une étude de Khaitovich *et al.* (Khaitovich *et al.*, 2006) avait émis l'hypothèse que si les lincRNAs n'étaient pas sous sélection, leurs diversités d'expression seraient les mêmes entre plusieurs tissus distincts (cerveau, cœur et tissu testiculaire) chez l'homme et

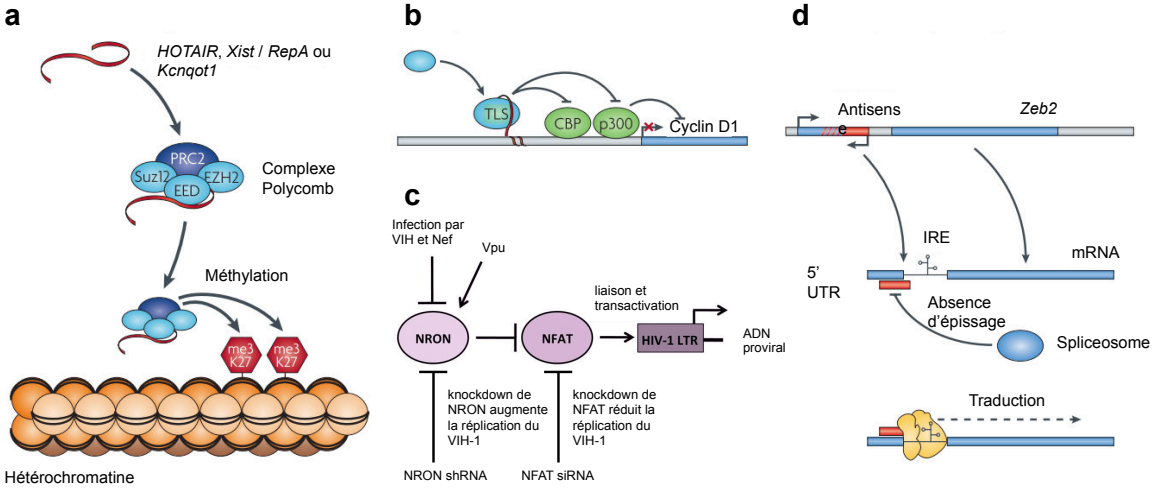

Figure 2 : Fonctions des longs ARN non-codants. *(a) Les lncRNAs peuvent recruter des complexes modificateurs de la chromatine à des locus génomiques spécifiques pour conférer leur activité catalytique. Dans ce cas, les lncRNAs HOTAIR, Xist, RepA ou Kcnqot1 recrutent le complexe Polycomb au locus HoxD, le chromosome X, ou le domaine Kcnq1, respectivement. Ils triméthylent le résidus 27 de lysine sur l'histone H3 (H3K27me3) pour induire la formation de l'hétérochromatine et réprimer l'expression des gènes. (b) Les lncRNAs peuvent réguler le processus transcriptionnel à travers une gamme de mécanismes. Les lncRNAs attachés au gène cycline D1 recrutent la protéine de liaison à l'ARN TLS pour moduler l'activité de l'histone acétyltransférase de la protéine de liaison CREB (CBP) et p300 pour réprimer la transcription des gènes. Traduit de Mercer et al., 2009. (c) Modèle proposé pour la régulation croisée de la réplication du VIH et du VIH-1. Le modèle montre l'infection par le VIH-1 et la protéine Nef virale pour réduire les niveaux de NRON, et la protéine Vpu virale pour augmenter les niveaux de NRON. Celui-ci à son tour régule l'activité du facteur de transcription NFAT, qui améliore la transcription du LTR du VIH-1. Les interventions de knockdown de NRON médiée par shRNA et la réduction de NFAT médiée par siRNA, sont montrées. Traduit de Imam et al., 2015. (d) Un lncRNA antisens peut masquer le site d'épissage en 5' du mRNA ZEB2 à partir du spliceosome, ce qui entraîne la rétention d'intron. La machinerie de traduction peut alors reconnaître et lier un site d'entrée de ribosome interne (IRE) dans l'intron retenu, ce qui entraîne une traduction et une expression efficaces de Zeb2. Traduit de Mercer et al., 2009.*

le chimpanzé. Suite à leurs analyses, ils ont observé, au contraire, que l'homme et le chimpanzé différaient entre ces tissus et que les transcrits intergéniques possédaient des rôles fonctionnels. Les lncRNAs présentent donc un grand nombre de rôles, ils interviennent au niveau épigénétique dans différentes modifications de la chromatine, au niveau pré-transcriptionnel et/ou post-transcriptionnel.

D'une part, les lncRNAs peuvent servir d'intermédiaire dans les changements épigénétiques en recrutant des complexes de remodelage de la chromatine spécifiques à différents loci. Ces lncRNAs présentent souvent des domaines de liaison à l'ARN mais peu de spécificité de séquence d'ADN, et sont capables de former des complexes ribonucléoprotéiques qui vont modifier (acétylation, méthylation…) la chromatine tout au long du développement d'une organisme (Mercer, Dinger and Mattick, 2009). Un premier exemple, *H19* est un lncRNA différentiellement exprimé entre les chromosomes de la mère et du père, la copie active étant présente chez la mère, et *H19* est donc sous empreinte génétique (ou 'imprinting'). Ce lncRNA est épissé, polyadénylé et exporté dans le cytoplasme où il s'accumule à des niveaux très élevés (Bartolomei, Zemel and Tilghman, 1991) et semble jouer un rôle dans la régulation de la croissance. Il a également été trouvé que le lncRNA *H19* est l'hôte d'un microRNA exonique, *miR-675* (Cai and Cullen, 2007). Historiquement, un des premiers lncRNAs caractérisés est donné par le lncRNA *Xist* ('*X Inactive Specific Transcript*') (Figure 2a.), dont le rôle est indispensable pour l'inactivation du chromosome X en *cis* chez les mammifères euthériens (Brockdorff *et al.*, 1991; Brown *et al.*, 1991). Ce lncRNA, d'une taille d'environ 17 kb, partage certaines similitudes avec *H19* car il est aussi épissé et polyadénylé. Cependant, *Xist*, comme la plupart des lncRNAs (Derrien, Johnson, *et al.*, 2012), échappe à l'exportation dans le cytoplasme et est plutôt associé sous forme de compartiment d'ARN avec le chromosome X qu'il inactive (Clemson *et al.*, 1996) via le recrutement de protéines de méthylation de l'ADN. Cette région de la chromatine qui est réduite au silence a fourni le premier modèle de lncRNA impliqué dans la stabilité épigénétique d'un gène *cis* silencieux.

D'autre part, les lncRNAs ont un rôle clé dans la régulation des processus de transcription. Les promoteurs proximaux peuvent être transcrits en lncRNAs, nous parlons de 'promoter associated ncRNAs' ou pancRNAs, qui recrutent et intègrent les fonctions des protéines de liaison à l'ARN dans le programme transcriptionnel. Ces lncRNAs peuvent, par

exemple, réguler l'activité de l'ARN polymérase II par interaction avec le complexe d'initiation pour influencer le choix du promoteur (Mercer, Dinger and Mattick, 2009). La transcription d'un lncRNA peut favoriser l'accessibilité de gènes codant pour des protéines à des ARN polymérases. De plus, les lncRNAs régulent également la transcription en *cis* indirectement, sans se lier à l'ADN (Ponting, Oliver and Reik, 2009), pour cibler des activateurs ou des répresseurs transcriptionnels via une variété de mécanismes. L'un de ces mécanismes utilise des signaux de dégradation de l'ADN pour induire l'expression de lncRNAs associés au promoteur du gène de la cycline D1 (Figure 2b.). Ces lncRNAs agissent en coopération pour moduler les activités de la protéine de liaison à l'ARN *TLS* ('*Transfer RNA-like structures'*). Cette dernière inhibe ensuite les activités d'histone acétyltransférase de la protéine de liaison CREB (CBP) et p300 pour réduire l'expression de la cycline D1.

Il a également été montré que les lncRNAs sont impliqués dans la régulation de l'expression des gènes codant pour des protéines et notamment des gènes qui résident dans la même région génomique sur un intervalle de quelques mégabases. L'expression des lncRNAs permet d'induire ou de réprimer l'expression du mRNA cible, mesurable par des corrélations statistiques, positives ou négatives, des profils d'expression (Signal, Gloss and Dinger, 2016). Cette approche corrélative parfois appelée 'Guilt-by-association' (coupable par association) permet l'étude de paires de lncRNAs et de mRNAs, notamment des lncRNAs capables de moduler l'expression d'un gène présent sur le même intervalle génomique, que nous définirons par *cis*-régulation par opposition à la *trans*-régulation qui désigne quant à elle, les séquences agissant à plus longue distance (par exemple, depuis un autre chromosome). Le principe de Guilt-by-association se base sur l'hypothèse que des gènes co-exprimés sont plus susceptibles d'être co-régulés, de partager des fonctions similaires, ou d'être impliqués dans des processus biologiques similaires. Ce principe, comme son nom l'indique, va attribuer des fonctions putatives aux gènes, basées sur les fonctions déjà connues des gènes co-exprimés (Signal, Gloss and Dinger, 2016).

D'autres lncRNAs régulent indirectement la transcription en contrôlant la localisation subcellulaire des facteurs de transcription. Ils agissent comme co-facteurs pour moduler l'activité de facteur de transcription, par exemple, ils peuvent affecter la localisation d'un facteur de transcription tel que *NFAT* (facteur nucléaire des cellules T activées) (Figure 2c.) par des interactions avec des facteurs de transport nucléaire (Willingham et al., 2005).

L'inactivation de ce lncRNA, appelé *NRON*, répresseur non-codant de *NFAT*, entraîne une augmentation du facteur de transcription *NFAT* dans le noyau et une augmentation de son activité.

Par ailleurs, des lncRNAs ont une capacité à reconnaître des séquences complémentaires permettant également des interactions hautement spécifiques qui sont susceptibles de réguler diverses étapes dans le traitement post-transcriptionnel des mRNAs, comme leur épissage, leur traduction, leur transport, ainsi que leur dégradation (Mercer, Dinger and Mattick, 2009). Les lncRNAs antisens peuvent masquer les *cis*-éléments clés pour le mRNA par la formation de duplex d'ARN, comme dans le cas du lncRNA antisens au mRNA *ZEB2* (*ZEB2-AS1*), qui peut masquer le site d'épissage en 5' du gène *ZEB2* ('*Zinc finger E-Box homeobox 2*') (Beltran *et al.*, 2008) à partir du spliceosome, entraînant une rétention d'intron. La machinerie de la traduction peut alors reconnaître et lier un site d'entrée de ribosome interne (IRE) dans l'intron retenu, permettant la traduction et l'expression de *ZEB2* (Figure 2d.).

Enfin, de plus en plus de lncRNAs sont mis en évidence dans le développement de maladies, tel que le lncRNA *HOXD-AS1* qui aurait un rôle de régulateur de l'apoptose cellulaire dans le carcinome humain (Lu *et al.*, 2017), ou encore le lncRNA *VELUCT*, qui, malgré sa faible abondance, apparaît comme essentiel à la viabilité des cellules de plusieurs lignées de cancers du poumon (Seiler *et al.*, 2017).

5. Implications des lncRNAs dans le développement des maladies et des cancers

Les petits ARN non-codants tels que les miRNAs spécifiques aux cancers ont largement été utilisés pour le diagnostic et le pronostic des cancers, car certains sont détectables dans le sang, l'urine, les expectorations et d'autres fluides biologiques des patients, leur pertinence en font de potentiels biomarqueurs (Weber *et al.*, 2010). Depuis l'apparition des nouvelles technologies de séquençage, les lncRNAs, présents également dans les fluides corporels, ont démontré leurs utilités comme marqueurs non invasifs en routine clinique (Ke *et al.*, 2017). Différentes études ont mis en évidence qu'un changement d'expression des lncRNAs était associé à la cancérogenèse, ainsi ces lncRNAs peuvent servir de cible pour le diagnostic et la thérapie de plusieurs cancers, comme montré dans l'étude sur les cancers gastriques (Ke *et al.*, 2017; Zhang *et al.*, 2018).

Un des lncRNA précurseur dans l'étude des tumeurs est le lncRNA *HOTAIR* identifié comme clé dans le diagnostic des tumeurs malignes du tractus gastro-intestinal (Liu *et al.*, 2015). Très largement étudié, il est aussi utilisé en tant que marqueur de pronostic défavorable lorsqu'il est positivement régulé dans les tumeurs mammaires primaires et métastatiques (Cai *et al.*, 2014; Serviss, Johnsson and Grandér, 2014; Kladi-Skandali *et al.*, 2015). *MALAT1* (*'Metastasis-Associated Lung Adenocarcinoma Transcript 1'*) est un autre lncRNA très étudié pour lequel sa sur-expression a été associée à l'état métastatique de tumeurs et son utilisation suggérée comme marqueur pronostique dans le cancer du poumon (Shi *et al.*, 2013). C'est maintenant de multiples fonctions oncogéniques qui sont associées aux lncRNAs dans le cancer du foie pour le lncRNA *HULC*, (*'Highly Up-regulated in Liver Cancer'*) (Yu *et al.*, 2017), et de la prostate (*PCA3*, *'Prostate Cancer Antigen'*) (Salameh *et al.*, 2015). Ces lncRNAs, *HULC* et *PCA3*, sont déjà utilisés pour le diagnostic du cancer du carcinome hépatocellulaire et du cancer de la prostate. Certains lncRNAs, tels que *LincRNA-p21* et *MEG3* (*'Maternally Expressed 3'*), ont été impliqués dans la modulation de la réponse du gène de suppression de tumeur *P53* (Zhou *et al.*, 2007; Yoon *et al.*, 2012). Dans certains cas, une combinaison de plusieurs lncRNAs est utilisée en tant que marqueur diagnostic ou pronostic comme par exemple chez des patients atteints de cancer gastrique, les quatre lncRNAs suivants sont détectés, *AK001058*, *INHBA-AS1*, *MIR4435-2HG*, et *CEBPA-AS1* (Ke *et al.*, 2017).

Une étude récente a aussi identifié un lncRNA appelé *SAMMSON*, exprimé spécifiquement dans les mélanomes cutanés chez l'homme, et qui jouerait un rôle majeur dans le développement de la tumeur. Le lncRNA *SAMMSON* est exprimé dans plus de 90 % des mélanomes humains et est une cible transcriptionnelle directe du facteur de transcription SOX10. C'est un biomarqueur candidat prometteur pour l'identification des lésions qui contiennent des cellules de mélanome malignes et invasives (Leucci, Vendramin, *et al.*, 2016). Toujours concernant le mélanome, les lncRNAs, tels que *SPRY4-IT1* (Khaitan *et al.*, 2011) et *BANCR* (R. Li *et al.*, 2014), ont également été validés comme dérégulés dans le mélanome et associés à la survie des patients atteints de mélanome cutané. Ces lncRNAs constituent une nouvelle niche pour le diagnostic et le pronostic ainsi qu'une nouvelle classe de cibles thérapeutiques cliniquement pertinentes pour le traitement et *in fine* une prise en charge du mélanome (Guo, Yao and Jiang, 2016; Leucci, Coe, *et al.*, 2016).

6. Conservation des lncRNAs entre espèces

Les preuves de l'importance des lncRNAs ne cessent de croître et de plus en plus de lncRNAs apparaissent comme associés à des maladies humaines (Batista and Chang, 2013; X. Li *et al.*, 2014; Li *et al.*, 2018; Raut and Khullar, 2018). Une identification exhaustive des lncRNAs conservés entre espèces, pourrait mettre en évidence des rôles biologiques conservés, permettre d'introduire les modèles animaux dans des essaies cliniques ou encore de développer des biomarqueurs pour différentes maladies.

L'identification des relations d'orthologie des lncRNAs n'est cependant pas facilité car le niveau de conservation de séquence est faible, probablement lié au fait que les lncRNAs ne subissent pas les mêmes contraintes sélectives que les mRNAs, ils ne sont en effet pas contraints par la traduction en acides aminés, ce qui permet une évolution des séquences des lncRNAs plus rapide que celle des mRNAs (Ulitsky, 2016). Une étude de génomique comparée est nécessaire afin d'observer la conservation des différents biotypes étudiés entre deux espèces. Pour optimiser cette analyse, deux approches complémentaires peuvent être utilisées et concernent l'analyse des séquences et l'étude de blocs de synténie.

L'analyse par comparaison de séquences est très largement utilisée pour mettre en évidence les relations d'orthologie des gènes codant pour des protéines entre espèces. Pour les lncRNAs orthologues, leur identification se fait à l'aide de séquences souvent plus courtes, et faiblement conservées, comme par exemple le lincRNA *MALAT1* ('*Metastasis-Associated Lung Adenocarcinoma Transcript 1'*), qui a été caractérisé comme régulateur de l'épissage alternatif de gènes cibles endogènes, et également en tant que gène favorisant la tumorigenèse (Li *et al.*, 2009). Dans l'étude d'Ulitsky *et al*., de 2011 (Ulitsky *et al.*, 2011), un orthologue de *MALAT1* caractérisé dans les cellules de mammifères (Tripathi *et al.*, 2010), est identifié chez le poisson-zèbre. Ils observent une similarité de séquence entre ce lincRNA chez le poisson-zèbre et les mammifères, limitée à l'extrémité 3' de la séquence. Malgré cette conservation limitée de séquence, la taille de *MALAT1* (7 kb) ainsi que l'absence d'intron efficacement épissé semblent être relativement fixé chez tous les vertébrés. Ces observations suggèrent que la fonctionnalité conservée de certains lincRNAs peut aller de pair avec un niveau relativement faible de séquence spécifique (Ulitsky et al. 2011). Même si des milliers de lincRNAs orthologues ont été reportés entre l'homme et la souris, certains possédant une similarité de séquence significative et reconnaissable entre

espèces (Ulitsky *et al.*, 2011), cette analyse demeure peu efficace pour déterminer les orthologues de lncRNAs entre différentes espèces.

En 1971, J.H. Renwick introduit le terme de synténie, étymologiquement 'sur le même brin', pour désigner des loci localisés sur le même chromosome (Renwick, 1971). En génomique comparative, ce terme décrit des régions chromosomiques conservées entre plusieurs espèces et plus généralement, des régions contiguës entre deux génomes, où l'ordre des gènes peut être altéré, ces régions seront définies comme segments conservés (CS pour 'Conserved Segment') ou blocs de synténie.

Ces approches à grande échelle et plus intégratives, améliorent la sensibilité de détection des orthologies des lncRNAs, notamment en combinant la similarité de séquence et la synténie. L'approche utilisée par la base de données Compara d'EnsEMBL (Herrero *et al.*, 2016), se base sur une heuristique qui évalue l'ordre des gènes en utilisant à la fois une mesure de similarité entre les gènes et d'alignement de séquence. Cette classe d'outils de détection d'orthologie tente donc d'inférer explicitement les phylogénies géniques et leur reconstruction dans différents arbres d'espèces. Cette approche a été adaptée pour supporter plusieurs génomes et étendue pour détecter les régions dupliquées. Les données sont analysées, et déjà disponibles sur la base publique d'EnsEMBL, et de fait, cette base permet son utilisation sans limitation pour une espèce déjà analysée. Par ailleurs, les données de blocs de synténies analysés et définis par Compara via les mRNAs, directement exploitables et améliorant la sensibilité de détection des orthologues de lncRNAs, permettent des prédictions plus fines que les approches purement basées sur la similarité de séquences.

La conservation de séquences de gènes orthologues entre espèces n'implique pas nécessairement une similarité du profil d'expression des gènes. La variation des profils d'expression entre gènes orthologues peut être liée à différents facteurs tels que la localisation tissulaire, le type cellulaire, ou encore le micro-environnement qui présente des caractéristiques propres dans chaque espèce. C'est grâce au développement des technologies d'analyse du transcriptome, telles que les puces d'expression et maintenant le RNA-seq, qu'émerge actuellement la transcriptomique comparative. Cette approche permet de comparer, avec des mesures quantitatives et qualitatives sans précédent, les profils et

niveaux d'expression et l'épissage des gènes au cours de l'évolution (Breschi, Gingeras and Guigó, 2017). Une application majeure de la transcriptomique comparative est de mettre en évidence les caractéristiques transcriptomiques récurrentes qui sont identifiées à travers les gènes orthologues (codant et non-codant), qui vont suggérer l'importance de l'expression, de la régulation et de la conservation évolutive. Cette analyse a toutefois pour limitation la considération des orthologues au préalable bien identifiés.

Une analyse comparée de l'expression des gènes humain et murin, à partir d'un cadre expérimental homogène, c'est-à-dire, des échantillons collectés, séquencés et analysés de manière similaire dans un même centre, a mis en évidence que globalement, les lncRNAs orthologues avaient des niveaux d'expression conservés (Breschi, Gingeras and Guigó, 2017). En outre, l'étude de Breschi *et al.* de 2017, a montré que la répartition de l'expression des lncRNAs au sein des différents tissus chez l'homme est conservée au sein de nombreux mammifères. Ainsi les lncRNAs exprimés de manière spécifique dans un tissu chez l'homme seront également tissu-spécifiques chez d'autres espèces et que les lncRNAs avec une distribution de type ubiquitaire chez l'homme seront ubiquitaires chez les autres espèces. Cependant ces résultats sont décrits pour les seuls 10 % de lncRNAs humains orthologues avec la souris et ne représentent qu'une faible proportion du catalogue humain (Breschi, Gingeras and Guigó, 2017).

7. Les éléments transposables et les lncRNAs

Plus récemment, les lncRNAs ont été caractérisés pour être particulièrement enrichis en éléments transposables (TEs), éléments mobiles qui ont largement contribué à l'évolution des génomes (Johnson and Guigó, 2014). Les TEs ont été caractérisés comme une composante majeure de nombreuses espèces végétales (Lisch, 2013; Hirsch and Springer, 2017) et également chez les mammifères (Alzohairy *et al.*, 2013). Ce sont des séquences d'ADN répétées qui sont de véritables fragments d'ADN mobiles capables de se déplacer de façon autonome dans le génome. Les TEs ont d'abord été découverts dans le maïs (McClintock, 1956), et mis en évidence pour leur rôle dans l'altération de la pigmentation du grains de maïs, avec l'insertion de TE dans un gène de la chaîne de biosynthèse des anthocyanes (Butelli *et al.*, 2012). De nombreuses insertions de TEs sont éliminées par la sélection naturelle car elles induisent des effets délétères. À l'inverse, des insertions de TE

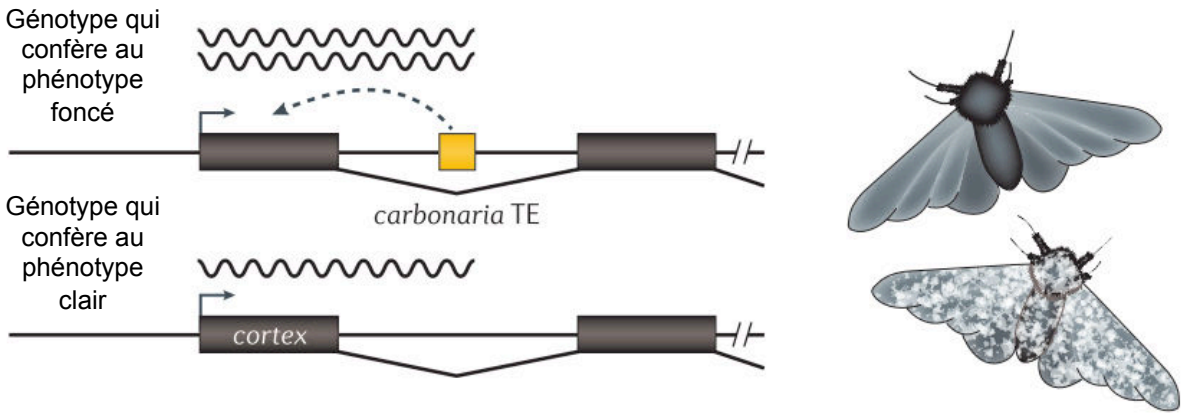

Figure 3 : Insertion du TE carbonaria dans un intron du gène du cortex chez le phalène du bouleau, *amélioration du niveau d'expression du gène cortex (la ligne pointillée indique un mécanisme de régulation non caractérisé), qui souligne la coloration adaptative qui s'est produite pendant la révolution industrielle. Traduit de Chuong et al., 2017.*

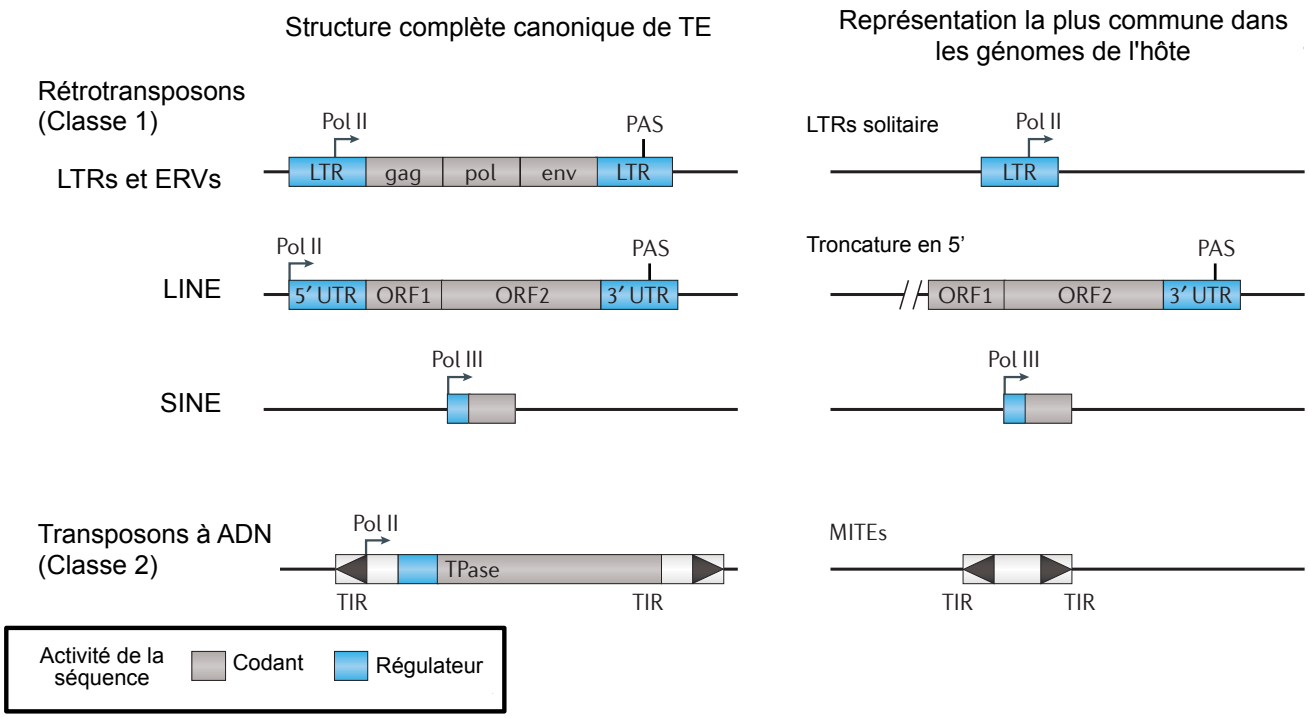

Figure 4 : Schéma des principales classes d'éléments transposables (TEs) et de leur organisation génétique typique*. Le panneau de gauche représente la version générale complète de chaque type de TE. La plupart des TEs hébergent des séquences régulatrices qui fonctionnent pour favoriser leur propre transcription et régulation, telles que les promoteurs de l'ARN polymérase II (Pol II) ou Pol III, et les signaux de polyadénylation (PAS). Le panneau de droite montre les structures de chaque type de TE telles qu'elles se produisent le plus souvent dans le génome, qui peuvent différer considérablement en fonction du TE. Traduit de Chuong et al., 2017.*

peuvent être avantageuses pour l'organisme, comme par exemple pour le phalène du bouleau (*Biston betularia*) où une insertion intronique d'un TE a été identifiée comme augmentant l'expression du gène *cortex* favorisant ainsi le phénotype plus foncé (Van't Hof *et al.*, 2016) (Figure 3). Dans cet exemple d'adaptation, les phalènes du bouleau n'ont pas été domestiqués en tant que tels, mais ont été soumis à une forte sélection artificielle due à la pollution par le charbon pendant la période industrielle (Chuong, Elde and Feschotte, 2017).

7.1. Les principales classes d'éléments transposables

Les TEs sont regroupés en deux grandes classes, d'une part, la classe I représentant les transposons à ADN et d'autre part, la classe II avec les rétrotransposons de type LTR pour 'long terminal repeat' ou non-LTR avec les 'long interspersed elements' (LINEs) et les 'short interspersed elements' (SINEs). Ces différentes classes d'éléments transposables sont retrouvées chez la majorité des mammifères.

Les LINEs représentent près de 20 % du génome humain, et leur insertion dans un exon peut fortement modifier le transcrit et le mécanisme d'épissage. Malgré un nombre important de mécanismes de défense de l'hôte mis en place pour restreindre la rétrotransposition de LINE-1, de la famille des LINEs, les insertions héréditaires LINE-1 continuent de se produire et contribuent à la variation humaine inter-individuelle (Richardson *et al.*, 2015). Bien que nous ne sachions toujours pas à quelle fréquence de nouveaux événements de rétrotransposition induite par LINE-1 peuvent survenir dans la population humaine, les estimations suggèrent que 1 % des hommes hébergent une nouvelle insertion (Richardson *et al.*, 2015).

Les TEs contribuent à une grande fraction des sites de liaisons aux facteurs de transcription à travers le génome (en moyenne ~ 20 %) (Chuong, Elde and Feschotte, 2017). Les LTRs représentent une plus grande proportion de ces derniers comparés aux autres TEs. De plus, le mécanisme de réplication de l'élément LTR engendre deux copies exactes du LTR de chaque côté des éléments codants dans le génome de l'hôte (Figure 4) (Chuong, Elde and Feschotte, 2017). Les LTRs sont en effet des séquences répétées en orientation directe qui contiennent les régions promotrices et régulatrices des séquences codantes de l'élément et qui comportent plusieurs domaines fonctionnels : U3, R et U5, désignées par rapport à leur

position sur le mRNA : U3 (pour unique en 3’ de l’ARN), R (pour répétée aux deux extrémités de l’ARN) et U5 (pour unique en 5’ de l’ARN). Les séquences promotrices ne sont actives qu’au niveau du LTR en 5’ ; les sites de polyadénylation ne sont actifs qu’au niveau du LTR en 3’. La région R des LTR se retrouve aux deux extrémités du transcrit et est indispensable pour la transcription inverse de ce mRNA en ADNc.

Un autre mécanisme, dit de ‘copier-coller’, permet aux SINEs de présenter un très grand nombre de copies (Richardson *et al.*, 2015). Par exemple, les éléments *Alu* sont les éléments SINEs les plus communs dans le génome humain, couvrant près de 10 % de la séquence du génome (Lubelsky and Ulitsky, 2018). Ces éléments *Alu* sont enrichis dans les régions transcrites et ont un impact important sur l'évolution du transcriptome. De tels événements sont activement supprimés dans les mRNAs, mais sont souvent retrouvés dans les lncRNAs (Kelley and Rinn, 2012).

Les éléments transposables sont source de variabilité génétique et présentent une importante valeur adaptative, comme par exemple, dans les mécanismes de défense aux stress biotiques et les mécanismes d’adaptation aux stress abiotiques. Des mutations peuvent être bénéfiques et sélectionnées par le génome hôte. Les protéines RAG1 et RAG2 seraient d’anciens TEs ‘domestiqués’ pour la formation des gènes d’anticorps par recombinaison (Kapitonov and Koonin, 2015). Rétrovirus et transposons sont à l’origine de la mise en place et de la diversification du placenta chez les mammifères (Cordonnier, Casella and Heidmann, 1995). Il a aussi été décrit que c’est l’infection par un rétrovirus et son intégration dans le génome qui aurait abouti à la mise en place du premier placenta (Villarreal, 2016). Enfin un dernier exemple a mis en évidence qu'un transposon de classe II (MER20), serait à l’origine de la régulation de l’expression des gènes localisés dans la paroi utérine (Emera and Wagner, 2012).

Alors qu'il y a de plus en plus de preuves qu'une grande partie des lncRNAs ont des fonctions multiples, notamment de régulation de l'expression des gènes codant pour les protéines, encore peu de connaissances sont établies sur les processus liés aux origines évolutives et à la diversification des lncRNAs. Ainsi des travaux ont déterminé que les TEs, de par leur capacité à introduire des séquences régulatrices lors de l'insertion chromosomique, constituent une force majeure dans le rôle des lncRNAs (Johnson and Guigó, 2014). Il a ainsi

été montré que certaines séquences de TE incluses dans les lncRNAs sont critiques pour leur biogenèse et probablement importantes pour leurs fonctions (Kapusta *et al.*, 2013). D'autres études, réalisées chez plusieurs mammifères ont aussi mis en évidence l'impact des TEs, notamment au cours du développement précoce. Ils peuvent également avoir des effets *cis*-régulateurs associés à des caractères tels que la taille ou la couleur chez les chiens ou les chats, ou encore à des risques de maladies (Chuong, Elde and Feschotte, 2017).

7.2. Un élément transposable canin spécifique

Les classes de TEs sont subdivisées en familles, et parmi la classe des SINEs, une famille de SINEC, pour SINE Carnivore, présente différents homologues décrits dans plusieurs espèces carnivores (Vassetzky and Kramerov, 2002). Il existe dans le génome canin un homologue le SINEC_Cf (Walters-Conte *et al.*, 2011), pour SINE Carnivore_*Canis familiaris,* le chien domestique. Les SINEC_Cf sont spécifiquement présents chez le chien, selon une étude qui, dès 2005 (Wang and Kirkness, 2005), avait mis en évidence que le génome canin contient une fréquence élevée d'allèles qui ne diffèrent que par la présence ou l'absence d'une répétition de SINEC_Cf révélant ainsi les coordonnées chromosomiques de plus de 10 000 loci bimorphiques pour les insertions de SINEC_Cf. Lorsque les SINEC_Cf sont transcrits dans l'orientation antisens, ils fournissent des sites accepteurs d'épissage, qui peuvent entraîner l'incorporation de nouveaux exons (Wang and Kirkness, 2005). De plus, la fréquence élevée des insertions de SINEC_Cf dans la population canine illustre un modèle de transcription allèle-spécifique qui peut être utile pour l'étude de l'expression génique différentielle entre races de chiens.

II. Le chien comme modèle d'étude

1. Les origines, du loup aux chiens modernes

Tous les chiens domestiques, *Canis lupus familiaris,* sont issus de la domestication du loup gris, *Canis lupus lupus*, datée entre 15 000 et 30 000 ans. Plusieurs scénarios sont proposées depuis quelques années, dont un qui montre, grâce à des analyses d'ADN moderne/actuel, d'une cinquantaine d'ADN ancien mitochondrial, un génome nucléaire ancien et aux analyses des archéo-zoologues, que à la fin de la période du Paléolithique supérieur, deux populations canines distinctes ont été domestiquées à l'Est et à l'Ouest de

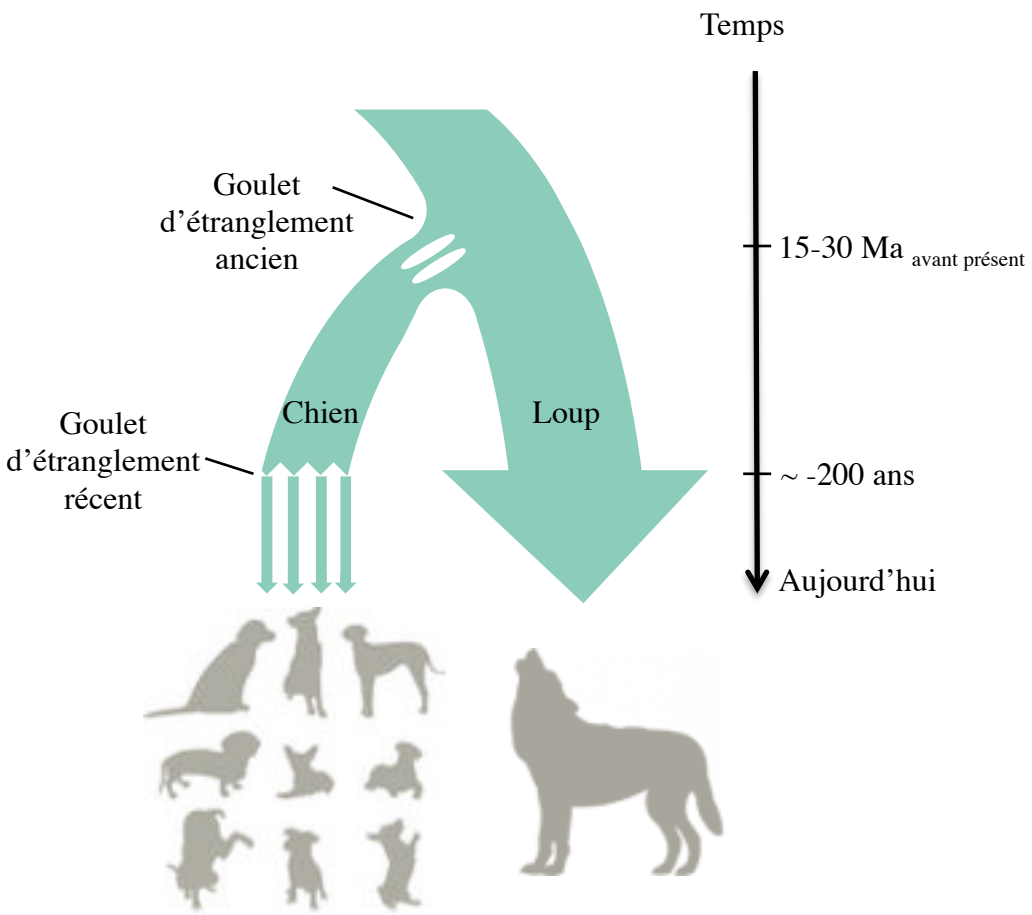

Figure 5 : Représentation schématique des deux goulets d'étranglement de l'histoire du chien. *Le premier correspond à la domestication du loup par l'homme puis, la création des races modernes. Source : adapté de Lindbald-Toh et al. 2005*

l'Eurasie (Frantz *et al.*, 2016). Les restes archéologiques des chiens domestiqués, datés de 15 000 ans avant notre ère sont caractérisés en Europe et il y a plus de 12 500 ans, c'est en Asie que, indépendamment, les hommes ont domestiqué les loups. Cette population d'Asie aurait migré d'Est en Ouest avec les populations humaines entre le 14ème et le 6ème millénaire avant notre ère et aurait peu à peu remplacé la population Européenne.

Un autre scénario est basé sur l'analyse de deux séquençages de génomes complets de prélèvements de chiens trouvés en Allemagne et datés du début et de la fin du Néolithique, il y a 47 000 ans et 7 000 ans. Les auteurs suggèrent une continuité de ces chiens jusqu'au chien moderne, sans identifier d'évènement de remplacement de la population comme proposé par Frantz *et al*., en 2016 (Frantz *et al.*, 2016; Botigué *et al.*, 2017).

Dans l'histoire de la domestication des premiers loups en chiens, nous observons deux goulets d'étranglement. Une première phase a séparé le loup du chien et a eu pour conséquence une forte réduction de la diversité génétique à partir d'un pool relativement restreint d'allèles (Lindblad-Toh *et al.*, 2005) (Figure 5). Les premières domestications ont engendré des changements morphologiques et comportementaux éloignant peu à peu le chien du loup.

À la suite des évènements de domestication, le chien a évolué avec l'homme pour s'adapter à un nouveau mode de vie. Cette adaptation est très bien illustrée par plusieurs études sur le gène *AMY2B* qui code pour une enzyme pancréatique, l'amylase, qui est essentielle dans la transformation de l'amidon en glucose au niveau du petit intestin. Deux études ont identifié une augmentation du nombre de copies de ce gène dans le génome du chien moderne allant de 4 à 34 copies selon les races canines contre 2 copies chez les loups (Arendt *et al.*, 2014; Ollivier *et al.*, 2016). Cette augmentation du nombre de copies de gènes ou CNV ('Copy Number Variation') est également associée à une augmentation de l'activité de l'enzyme amylase 2b (Arendt *et al.*, 2014). Plus récemment, l'analyse de 13 ADN anciens issus de 8 sites en France, Suisse, Roumanie, et Turkmenistan (Ollivier *et al.*, 2016) a permis de montrer que cette acquisition de copies surnuméraires d'*AMY2B* a eu lieu dès le 7ème millénaire avant notre ère pour le site en Roumanie (8 copies), et est également identifiée en France et Turkmenistan (-4 000 et -5 000 ans), période du Néolithique. Cette étude permet d'écarter une origine moderne de l'augmentation du nombre de copie, et de

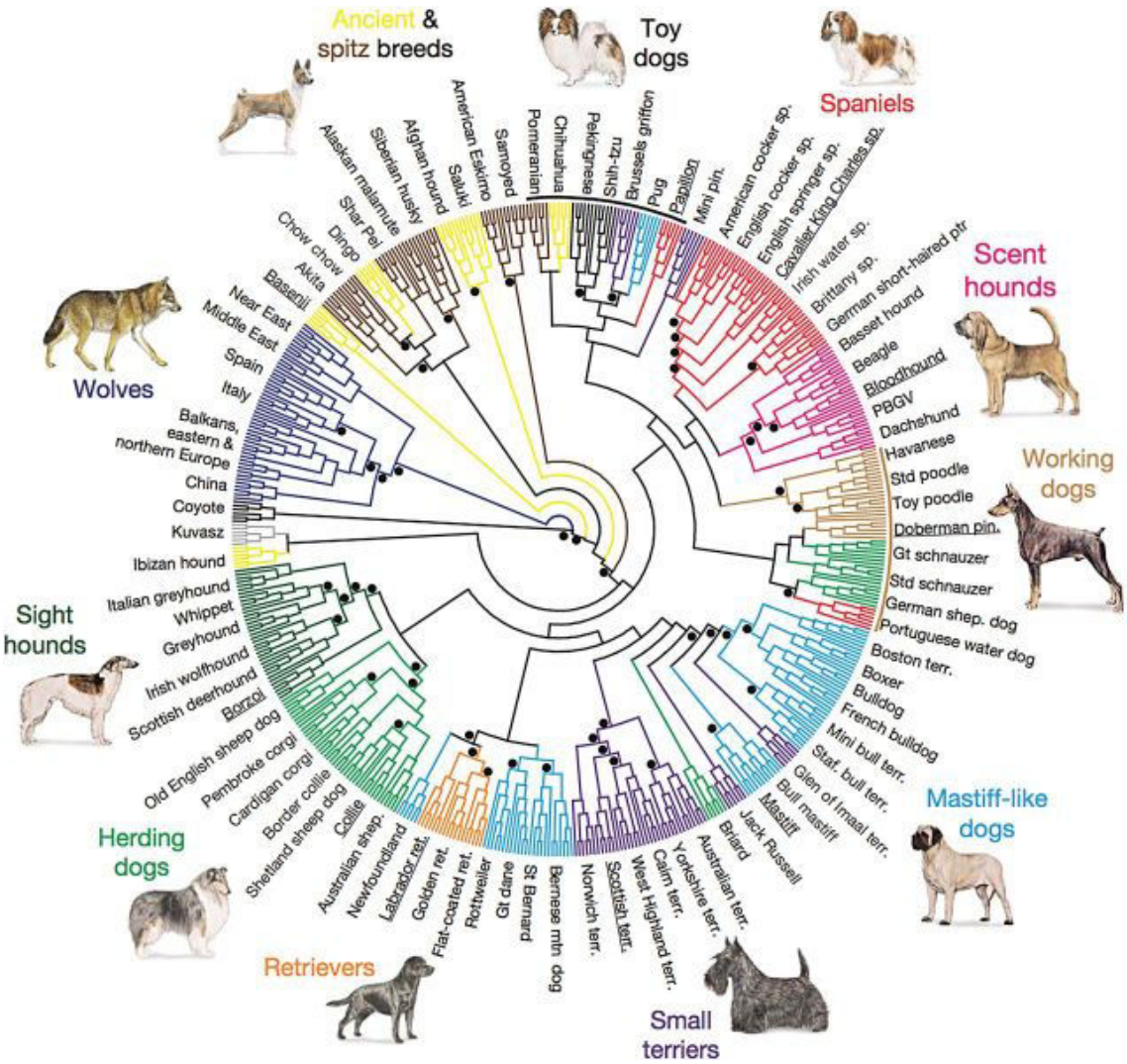

Figure 6 : Phylogénie des races de chiens. *Les comparaisons de 170 000 marqueurs SNPs analysés par fenêtre de 10 SNPs ont permis d'identifier des régions chromosomiques propres à chaque race. Chaque couleur de branche correspond à une fonction particulière du chien (garde, chasse...). D'après Vonholdt et al., 2010.*

montrer que des millénaires de domestication du chien ont conduit à de profondes modifications de son génome. L'étude illustre ainsi une adaptation bio-culturelle entre l'homme et le chien, avec l'adaptation du chien au mode de vie de l'homme à cette période avec l'accès à de nouvelles sources de nourriture par rapport au loup et le rôle de l'agriculture sur la sélection du génome canin (Ollivier *et al.*, 2016).

Au cours des siècles, l'homme a créé des populations puis des races canines par sélection plus ou moins intenses, qui ont entrainé un deuxième goulet d'étranglement génétique (Lindblad-Toh *et al.*, 2005) (Figure 5). Dans un premier temps, les chiens ont aidé à la chasse, puis à la protection du territoire ou encore des troupeaux. Ensuite l'homme a sélectionné des critères comportementaux, pour la compagnie. Quelques races (Labradors, Golden Retrievers) ont également été sélectionnées pour des chiens d'assistance aux personnes handicapées, pour leurs bonnes capacités d'apprentissage. D'autres sélections ont été réalisées sur des critères physiques et esthétiques, développés dans les clubs de races.

Cette sélection a entraîné des différences morphologiques entre races canines dépassant largement celles existantes au sein de tout autres espèces mammifères. Par exemple, un Chihuahua de 1 kg et 20 cm de haut et un Terre-neuve de 60 kg et 80 cm de haut représentent des morphotypes extrêmes. Ces deux races, appartiennent pourtant à la même espèce mais diffèrent d'un facteur 60 pour leur poids et d'un facteur 4 pour leur taille. Les différences de taille et de poids sont les plus évidentes, mais les races se distinguent sur de nombreux autres traits morphologiques comme la forme du crâne, des oreilles, du museau, de la queue, des variations sur la nature des poils (longs, courts, frisés ou absents) ou leurs couleurs. La plupart de ces races sont très différentes entre elles et sont réparties en 10 groupes par la Fédération Cynologique Internationale (FCI) (Figure 6). Cette classification est principalement basée sur la morphologie et les aptitudes des différentes races sans forcément tenir compte de l'historique de leur création. Par exemple, le groupe 1 regroupe tous les chiens de bergers, le groupe 3 rassemble les terriers et le groupe 10 correspond aux lévriers. D'autres groupes présentent des phénotypes beaucoup plus variés. Ainsi, nous retrouvons dans le groupe 2, des chiens ayant des tailles bien différentes, dont le Pinscher Nain mesurant environ 25 cm au garrot pour 5 kg, le Dogue de Bordeaux mesurant

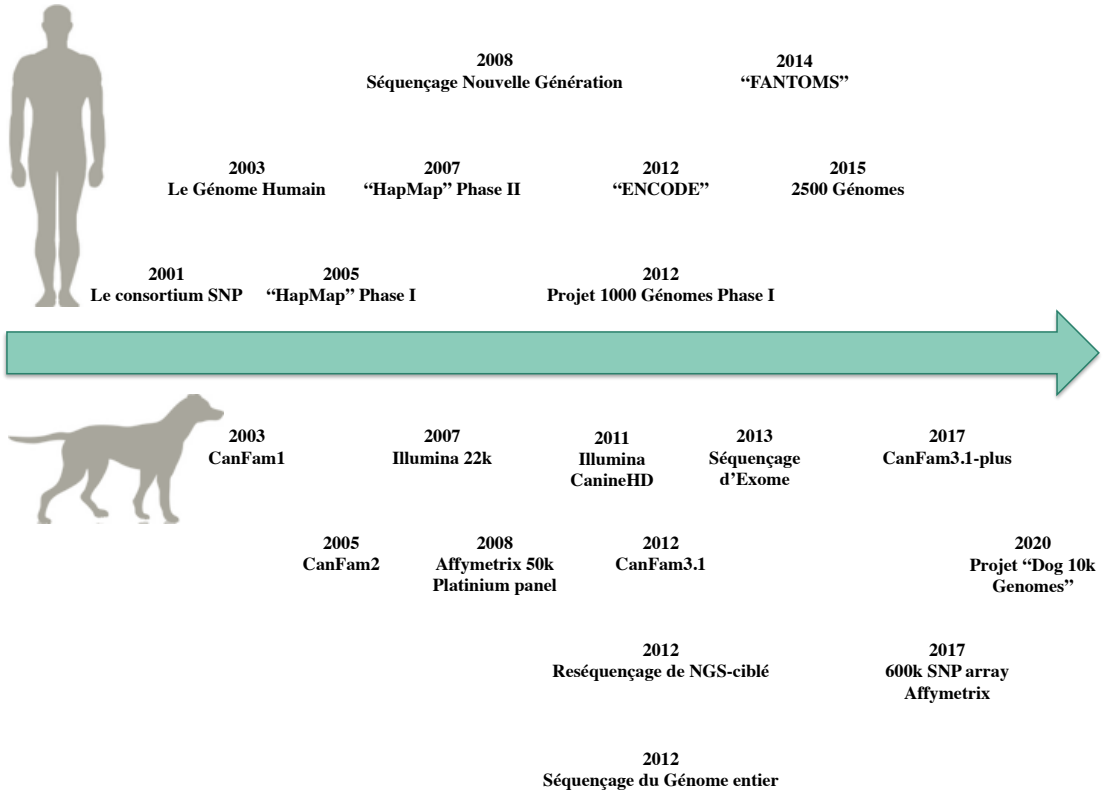

Figure 7 : Avancée de la génétique humaine et canine. *Les différentes étapes, projets et consortia du développement d'outils génétiques chez le chien et l'homme. CanFam1, CanFam2 et CanFam3.1 correspondent aux différentes versions d'assemblage du génome canin. Source : adapté et traduit de Steenbeek et al., 2016.*

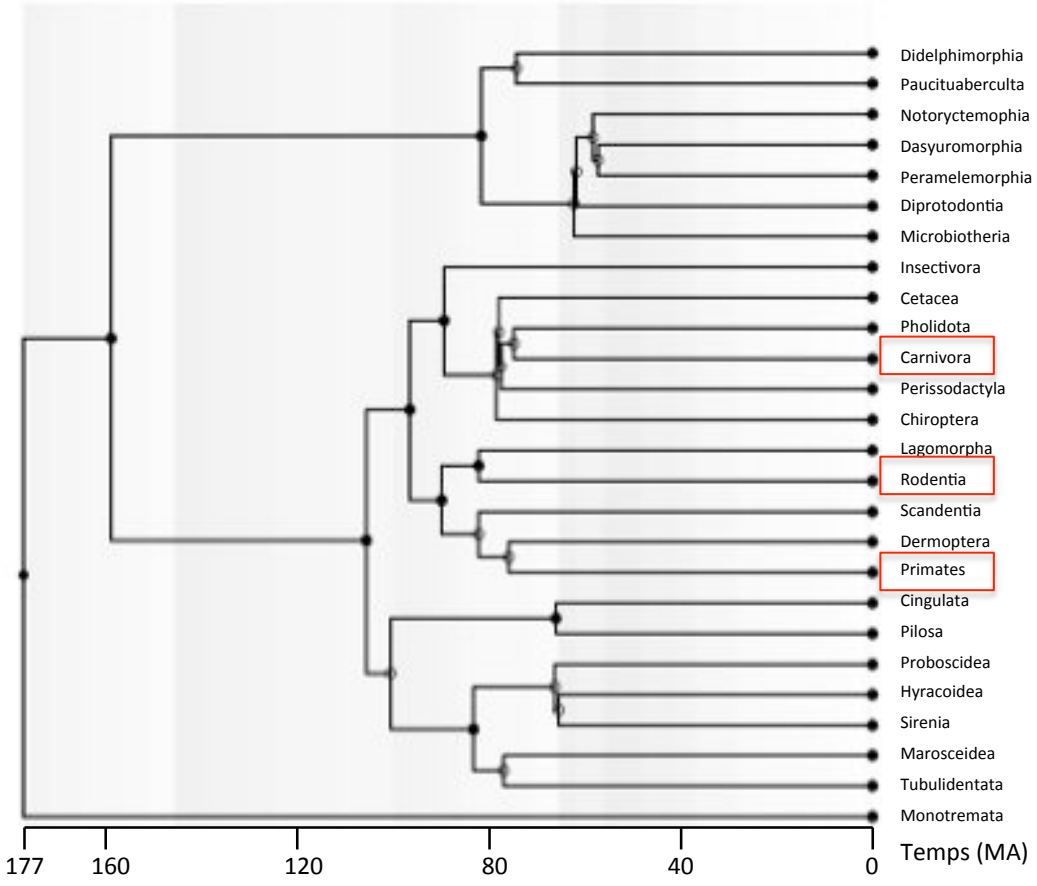

Figure 8 : Arbre phylogénétique. *Les espèces encadrée en rouge correspondent aux primates, aux carnivores et aux rongeurs, espèces modèles discutées dans le manuscrit.*

65 cm pour 60 kg, le Dogue Allemand, 85 cm pour 60 kg ou encore le chien de Montagne des Pyrénées, 75 cm pour 55 kg.

2. L'annotation du génome canin

Le chien domestique, *Canis lupus familiaris*, est le 5ème organisme à avoir été séquencé entièrement par la méthodologie du séquençage Sanger (Figure 7) (Lindblad-Toh *et al.*, 2005). Dès 2003, le génome d'un caniche mâle avait été séquencé avec une profondeur relativement faible de 1,5 X, ce qui signifie que chaque nucléotide a été séquencé en moyenne 1,5 fois, constituant une première version du génome canin appelée CanFam1 (*Canis Familiaris* 1) (Kirkness *et al.*, 2003). Ce séquençage léger a permis plusieurs études d'annotation fonctionnelle, structurelle et de génomique comparative du génome canin, notamment la confirmation que la séparation des carnivores a été plus précoce que celle des rongeurs (~ 95 MA versus 90 MA) (Figure 8) et l'identification, via les cartes hybride d'irradiation (RH), de blocs de synténie conservés par la construction des cartes comparées chien/homme très denses (Hitte *et al.*, 2005). Ce séquençage a révélé qu'au niveau nucléotidique, le génome canin présente un taux de substitution comparable à celui de l'homme alors qu'il est 1,6 fois supérieur entre le chien et la souris. De plus, environ 25 % de séquences non-répétées du génome canin s'alignent à cette profondeur avec le génome humain, valeur à mettre en correspondance avec les 40 % de séquences murines issues d'un séquençage profond s'alignant avec le génome humain. Enfin, ce projet a mis en évidence près de 18 500 fragments de gènes canins orthologues avec les gènes humains.

Deux ans plus tard, en 2005, un projet international piloté par l'actuel BROAD Institute de Cambridge (USA) a réalisé le séquençage d'une femelle boxer. Lors de ce séquençage 35 millions de séquences ont été produites couvrant 99 % des 2,4 Gb du génome canin, avec une profondeur plus importante de 7,5 X, (Lindblad-Toh *et al.*, 2005). Ce séquençage a permis d'obtenir un nouvel assemblage et une nouvelle annotation du génome canin, CanFam2. Lors du projet en 2005, 60 races et 120 chiens différents ont été étudiés et le choix s'est arrêté sur la race du boxer basée sur des travaux préalables (Parker *et al.*, 2004) qui avaient démontré un faible taux d'hétérozygotie chez cette race. Cette faible variabilité génétique est considérée comme avantageuse pour la phase d'assemblage. L'intérêt de

Types / Biotypes		CanFam3.1	CanFam3.1-plus	Nouveau
Gènes	mRNAs*	21 474	21 810	336
	lncRNAs**	8 008	10 444	2 436
Transcrits	mRNAs	100 110	158 750	58 640
	lncRNAs	12 506	22 880	10 374

Table 1 : Comparaison des annotations CanFam3.1 versus CanFam3.1-plus*. Les valeurs en gras correspondent aux nombres de gènes et transcrits nouveaux de CanFam3.1-plus. * Les gènes avec au moins un transcrit mRNA ; ** les gènes avec au moins un transcrit lncRNA (sans aucun mRNA). Table extraite et traduite des 'Supplementary' de Wucher et al., 2017.*

travailler à partir d'un génome femelle permet de disposer de deux copies du chromosome X, au détriment du séquençage du chromosome Y.

Actuellement les approches de séquençage haut-débit permettent non seulement l'étude des gènes codant pour des protéines, mais aussi des longs ARN non-codants (lncRNAs), des transcrits antisens, des microRNA (Hoeppner *et al.*, 2014; Broeckx *et al.*, 2015; Penso-Dolfin *et al.*, 2016), des épissages alternatifs, des pseudogènes, des régions régulatrices et des insertions/délétions (Derrien, Vaysse, *et al.*, 2012). Un nouvel assemblage et une nouvelle version de l'annotation du génome canin, CanFam3.1 (Hoeppner *et al.*, 2014), est présentée en 2014 et permet de combler des milliers de gaps et d'améliorer l'annotation des gènes codant pour les protéine (21 474 mRNAs) et de proposer un premier catalogue de lncRNAs canins (8 008 lncRNAs) (Table 1). Plus récemment, une extension de l'annotation du génome appelée CanFam3.1-plus, a été réalisée dans notre équipe et inclus l'annotation de 336 nouveaux mRNAs et de 2 436 nouveaux lncRNAs permettant de porter à 10 444 le nombre de gènes lncRNAs et de mieux explorer leurs implications dans les mécanismes de régulation de l'expression génétique chez le chien (Wucher *et al.*, 2017). Très peu de lncRNAs orthologues ont pu être mis en évidence entre le chien et l'homme, contrairement aux mRNAs où la base de données Ensembl Compara permet d'identifier 15 448 orthologues 1:1 qui possèdent en moyenne 84 % d'identité de séquence. À titre de comparaison, la souris et l'homme présentent 15 926 gènes codant pour des protéines orthologues 1:1 avec 83 % d'identité de séquence en moyenne.

Enfin, la communauté scientifique canine s'organise actuellement pour réaliser le séquençage de génomes de 10 000 canidés. En effet, le consortium 'Dog 10K genomes project'[7] inspiré du projet de séquençage chez l'homme 'The 1000 Genomes Project Consortium' (1000 Genomes Project Consortium *et al.*, 2015) a été mis en place par les Drs Elaine Ostrander (NIH, USA), Robert Wayne (UCLA, USA) et Ya-Ping Zhang (Kunming Institute of Zoology, Chine). Ce projet ayant débuté par le séquençage de 1 000 génomes canins au printemps 2016, prévoit de construire une ressource complète pour la communauté canine d'ici 2020. Les principaux objectifs étant d'améliorer la compréhension de l'histoire évolutive

[7] http://dog10kgenomes.org

Figure 9 : La diversité des races de chiens*. Les races varient en fonction de nombreux traits, y compris la taille, la longueur des pattes, le pelage, la couleur et la forme du crâne. Sont représentés : Barzoï (A), Basset Hound (B), Chihuahua (C), Schnauzer Géant (D), Bichon frisé (E), Colley (F), Bouledogue Français (G), Teckel (H), Braque Allemand à poil court (I), Épagneul Nain continental papillon (J) et Mâtin Napolitain (K). Source : Ostrander 2012.*

du chien et des changements génétiques associés à la domestication, ainsi que de créer une base de donnée de variants exhaustive. Les informations sur le génome canin s'étoffant, il sera plus aisé de pointer les variations génétiques responsables notamment de maladies.

3. Les prédispositions du chien à différentes maladies génétiques

La FCI reconnaît 344 races de chien, témoignant d'une très grande diversité phénotypique et comportementale unique et inégalée chez les mammifères (Figure 9). La sélection de critères particuliers et une forte utilisation en reproduction de mâles correspondants parfaitement aux critères de sélection, dit 'étalons champions', a engendré des descendances consanguines et une perte d'hétérozygotie dans les races définies par la FCI et la SCC (Société Centrale Canine) en France. Les races de chiens ont donc été créées par des pratiques de sélections artificielles sévères conduisant à une forte homogénéité génétique intra-race et une hétérogénéité inter-race. Une race de chien peut donc être qualifiée d'isolat génétique. Par comparaison, la variation génétique entre les races représente plus de 27 % de la variation observée dans toute la population canine, alors que la différenciation génétique reportée pour les populations humaines est de 5 à 10 % (Parker *et al.*, 2004).

L'espèce canine est caractérisée par une structure génétique toute à fait particulière, qui se traduit par une forte perte de variabilité génétique au sein des races. Ce véritable goulet d'étranglement de la diversité génétique se traduit sur le plan de la structure génétique par un fort déséquilibre de liaison (DL) qui réfère à la tendance des deux loci sur le même chromosome à être hérités ensemble. Bien que nous pourrions penser que la population de chiens domestiques soit très variable et donc susceptible d'avoir de nombreux événements de recombinaison et de faibles taux de DL, des études antérieures ont montré que l'étendue du DL était plus longue chez les chiens de race (Sutter *et al.*, 2004; Ke *et al.*, 2011). Les chiens de race pure présentent en effet des signes de goulets d'étranglement génétiques et une incidence élevée d'effets fondateurs qui entraînent un DL plus étendu que celui auquel nous pouvons s'attendre. Le fait que les populations fondatrices soient petites et que les pratiques d'élevage aient parfois eu recours à la consanguinité a considérablement accentué ce mécanisme. Ces régions en DL de grande taille et la forte homozygotie des blocs haplotypiques font de ces races un 'outil' de choix pour les analyses

génétiques basées sur la liaison ou l'association génétique. En effet, dans une population donnée, des DL longs permettent d'utiliser un plus petit nombre de marqueurs pour interroger le génome, et également moins d'échantillons pour une puissance statistique suffisante. Les études d'associations pangénomique (GWAS, 'Genome-Wide Association Study') chez le chien sont grandement facilitées par rapport à l'homme (Hayward *et al.*, 2016). La découverte de gènes associés ou responsables de maladies est a priori plus aisée au sein des races de chien que sur l'ensemble de la population humaine car les individus atteints, partageant des ancêtres communs, ont une probabilité plus élevée de partager les mêmes causes génétiques.

Les pratiques de sélections artificielles ont donc entrainé de nombreuses prédispositions aux maladies génétiques dans beaucoup de races canines. De plus, les maladies génétiques ont des prévalences bien supérieures à celles observées chez l'homme. Une maladie est considérée chez l'homme comme très fréquente à partir d'un individu atteint sur 500 (0,2 %), or il n'est pas rare de voir au sein des races de chien des fréquences de maladies allant de 1 à 10 % (Giger, Sargan and McNiel, 2006), voir même jusqu'à 30 % comme dans le cas de l'ichtyose du Golden Retriever, une maladie qui se manifeste par l'apparition de pellicules et peut déclencher des infections de la peau du chien (Grall *et al.*, 2012).

Le chien partage depuis des millénaires la vie et le même environnement que l'homme et est exposé aux mêmes facteurs environnementaux. De ce fait, il représente un modèle très adapté pour la plupart des études de recherche des causes génétiques des cancers ou des maladies multifactorielles impliquant le milieu environnemental, en comparaison d'autres modèles, comme la souris par exemple. De plus, compte tenu de sa taille et de son métabolisme plus proche de l'homme que celui de la souris, le chien est déjà utilisé en routine pour des tests pharmacologiques depuis les années 2000 (Starkey *et al.*, 2005). L'importance du chien dans notre société a fait qu'il possède un suivi médical important. Les techniques de médecines vétérinaires se sont améliorées durant les dernières décennies et nous permettent de répertorier avec soin plusieurs maladies chez le chien. Selon la base de

donnée OMIA, 'Online Mendelian Inheritance in Animals' [8], près de 310 maladies Mendélienne sont dénombrées chez le chien. Actuellement la SCC contient plus de 227 000 chiens inscrits au LOF (Livre des Origines Français) en 2016 qui fait de l'espèce canine un vaste 'réservoir' de maladies génétiques dont beaucoup de ces maladies sont homologues à celles décrites chez l'homme.

Aujourd'hui, la plupart des races canines sont atteintes de diverses maladies génétiques naturelles et spontanées. S'intéresser aux causes génétiques des maladies canines est important à la fois en médecine vétérinaire et en tant que modèles pour les maladies humaines. Chez l'homme, les maladies génétiques sont souvent complexes et multifactorielles, rendant difficile la recherche des évènements majeurs de la prédisposition et des causes génétiques. Chez le chien, les races sont plus homogènes sur le plan génétique simplifiant ainsi les analyses visant à identifier les gènes impliqués dans certaines maladies. Pour une malade complexe, une race présentera un plus petit nombre d'altérations génétiques et/ou de gènes impliqués dans la maladie et une autre race pourra présenter des causes génétiques différentes, ce qui n'est pas le cas chez l'homme, présentant une multitude de gènes du fait de la plus forte hétérogénéité des populations considérées lors des études génétiques. Pour faciliter l'étude de ces maladies, la souris est souvent utilisée en laboratoire en induisant des mutations qui vont mimer les maladies étudiées. Le chien est un modèle génétique d'intérêt pour l'étude des maladies homologues aux maladies humaines car il partage le même environnement que l'homme, ainsi que leur physiologie. De plus, la survenue des maladies se fait de manière spontanée et les possibilités d'études familiales et populationnelles liées à la disponibilité de larges fratries et de vastes cohortes, renforcent l'intérêt de ce modèle. En effet il n'est pas rare de compter dans une même portée près de dix chiots et compte tenu de l'utilisation d'étalons champions et des pratiques d'élevage, nous pouvons établir les relations familiales et génétiques pour un grand nombre d'individus et de nombreux animaux malades. De plus, les chiens de pure race sont inscrits dans les registres du LOF (Livre des Origines Français) qui est le registre d'inscription national des chiens de race, qui permettent un accès aux pédigrées et aux informations généalogiques, via les clubs de race et les sociétés canines nationales.

---

[8] http://omia.org/home/

Sous-type de cancer	Race de chien
Lymphome (non spécifié) - Lymphome B - Lymphome T	Bobtail, Boxer, Pointer, Golden Retriever, Rottweiler, St-Bernard, Scottish Terrier, Bouledogue - Lévrier Irlandais, Husky Sibérien, Shih Tzu, Airedale Terrier, Cavalier King Charles Spaniel, Yorkshire Terrier - Boxer, Cocker, Basset Hound
Ostéosarcome	Grandes et très grandes races, Lévrier Irlandais, Iévrier Écossais, Dogue Allemand, Bouvier Bernois, Mastiff, St-Bernard, Setter Irlandais, Golden Retriever, Rottweiler, Dobermann, Lévrier
Sarcome des tissus mous	Grands chiens, comme Boxer, Bouvier Bernois, Airedale Terrier, Dogue Allemand, St-Bernard, Basset Hound, Golden Retriever – tous avec des récurrences deux fois plus grandes que dans la population générale canine
Hémangiosarcome	Berger Allemand, Bouvier Bernois, Golden Retriever, Retriever à poil plat, chien d'eau Portugais, Labrador du Retriever, Boxer, Skye Serrier, Berger Australien
Sarcome histiocytaire	Bouvier Bernois, Retriever à poil plat, Rottweiler, Golden Retriever
Tumeur des mastocytes	Boxer, Carlin, Labrador du Retriever, Golden Retriever, Braque Hongrois
Méningiomes	Races mésocéphales et dolichocéphales (respectivement au nez moyen ou long), par exemple Labrador du Retriever, Golden Retriever, les Colleys
Gliomes (incluant glioblastome)	Races brachycéphales (au nez court), y compris les Boxers, Bouledogues et Terriers
Séminome testiculaire	Chien d'élan Norvégien
Carcinome de la cavité nasale	Golden Retriever, Beagle, Boston Terrier, Colley à poil long, Berger Belge
Carcinome de la vessie	Scottish Terrier, Beagle, West Highland White Terrier, Berger des Shetland, chien Esquimau Américain, Schnauzer Standard
Carcinome du tractus urinaire	Airedale Terrier, Beagle
Carcinome squameux	Caniche Standard, Schnauzer Géant
Mélanome - Mélanome buccal - Mélanome cutané	- Caniches - Schnauzers, Berger de Beauce

Table 2 : Exemples de cancers associés à des races de chien spécifiques. *Traduit de l'article de Schiffman et al., 2015*

Certaines races de chien sont prédisposées à des maladies en particulier ; notamment des maladies rares. Nous avons observé une prédisposition de l'atrophie progressive de la rétine, APR, dans deux races, le Border Collie et le Berger Picard, cette maladie est similaire aux rétinites pigmentaires chez l'homme (Vilboux *et al.*, 2008). Nous retrouvons aussi un symptôme d'automutilation souvent fréquent dans les races de chien de chasse. Une étude sur une neuropathie, homologue aux neuropathies héréditaires sensitives et autonomes (HSANs) chez l'homme, a été réalisée et a permis l'identification d'une mutation ponctuelle en amont du gène *GDNF* et située dans le dernier exon d'un ARN long non-codant, *GDNF-AS* (Plassais *et al.*, 2016). De nombreuses de races sont également prédisposées à différents cancers (Schiffman and Breen, 2015) (Table 2), par exemple les races brachycéphales, caractérisées par des crânes plus larges que profonds, sont sujettes aux gliomes, avec des fréquences bien plus importantes que dans les autres races (Thomas *et al.*, 2009). Une forte prédisposition aux cancers est également reportée dans de nombreuses races, telles que la race Caniche qui présente une forte prédisposition aux mélanomes de type non-UV dépendant comme les mélanomes oraux (Gillard *et al.*, 2014).

III. Le chien, modèle spontané des cancers de l'homme, exemple du mélanome muqueux

1. L'intérêt d'un modèle spontané

Les cancers présentent une incidence en augmentation. En France, le nombre de nouveaux cas est passé de 155 004 en 2012 à 399 500 en 2017[9]. Un manque de moyens thérapeutiques est observé pour le traitement de certaines tumeurs, il s'agit donc d'une préoccupation majeure de santé publique. Les multiples présentations cliniques des cancers et la complexité de leur étiologie sont des obstacles majeurs auxquels doit faire face la communauté scientifique pour mieux comprendre les causes et la progression de cette maladie multifactorielle. Face à ces difficultés, les races canines, qui développent spontanément les cancers représentent une ressource pour l'étude de ces cancers dans un organisme physiologiquement proche de l'homme. En effet, les cancers chez le chien

[9] http://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers

regroupent toutes les spécificités qui font l'intérêt de ce modèle pour l'étude de maladies complexes. Ils présentent un développement naturel, dans des races fortement prédisposées, de cancers cliniquement homologues à ceux de l'homme. Chez l'homme, la grande variabilité du fond génétique restreint le succès des études d'association génétique qui montrent à présent leurs limites. En effet, après l'identification de déterminants majeurs du risque de développer des cancers, ces travaux nécessitent aujourd'hui un nombre plus élevé de cas et de contrôle pour identifier des effets génétiques souvent faibles. Contrairement à l'homme, le chien présente une structure génétique singulière dans la population (Kirkness *et al.*, 2003; Lindblad-Toh *et al.*, 2005), elle offre une grande puissance statistique pour la cartographie de loci associés à des maladies et permet la découverte d'altérations génétiques prédisposantes. Cette prédisposition raciale spécifique témoigne de l'opportunité unique qu'offre le chien pour identifier des causes génétiques spécifiques des différents types de cancers.

2. L'oncologie comparée entre le chien et l'homme

Ces dernières années, dans les pays riches, l'espérance de vie des chiens de compagnie a augmenté notamment grâce à une meilleure prise en charge vétérinaire, à l'investissement croissant moral et financier des propriétaires dans la santé animale et au développement d'assurances de santé pour chiens et chats (Paoloni and Khanna, 2008). Malheureusement, les chiens de plus de 10 ans présentent de plus en plus de cancers, notamment chez les chiens de pure race (Sutter and Ostrander, 2004), 50 % d'entre eux développent la maladie et environ 25 % de ces derniers en meurent. Les cancers touchent également des jeunes chiens avec des causes génétiques et environnementales tout comme l'homme (Bronson, 1982; Adams *et al.*, 2010; Dobson, 2013; Davis and Ostrander, 2014).

Pour un chien qui vit en moyenne 12 ans et un homme 79 ans, le vieillissement est six à huit fois accéléré chez le chien et l'évolution tumorale est également plus rapide et plus agressive, dû en partie par le fait que le diagnostic soit posé plus tard dans l'avancement de la maladie, ce qui offre la possibilité de répondre à de nombreuses questions sur la progression du cancer, dans une période beaucoup plus courte que lors d'essais menés chez l'homme. Par exemple, les intervalles sans maladie ou encore appelé survie sans rechute dans l'étude de Paolini et Khanna de 2008 (Paoloni and Khanna, 2008), sur l'ostéosarcome et

le lymphome canin, ont été évalués en seulement 18 mois, alors que l'étude de paramètres similaires sur les mêmes cancers chez l'homme ont exigé plus de 7 ans (MacDonald *et al.*, 2005; Leahy *et al.*, 2006). À cela, s'ajoute le fait qu'ils soient exposés aux mêmes facteurs de risques environnementaux. De plus, les cancers canins apparaissent spontanément et sont cliniquement, histologiquement et dans leur réponse aux traitements, très semblables aux cancers humains. Les récentes études moléculaires sur ces cancers canins montrent qu'ils présentent les mêmes altérations génétiques et moléculaires (Schiffman and Breen, 2015). L'analyse moléculaire de ces cancers chez le chien a déjà permis d'identifier de nombreuses altérations géniques, sur des oncogènes et des gènes suppresseurs de tumeurs, identiques chez l'homme et le chien (Ulvé *et al.*, 2017). Le chien peut donc aider l'homme à découvrir de nouvelles voies oncogéniques ou à mieux déchiffrer les voies déjà connues chez l'homme et à les cibler par différents essais thérapeutiques.

Enfin, nous remarquons qu'à l'instar de nombreuses autres maladies, il existe certains sous-types de cancers spécifiques d'une ou d'un petit nombre de races, présentant des prédispositions parfois très importantes. Entre autres exemples, les grandes races (Dogue Allemand, Saint-Bernard, Leonberg…) ont d'avantage tendance à développer des ostéosarcomes (Ru, Terracini and Glickman, 1998; Rowell, McCarthy and Alvarez, 2011). Le Boxer, le Golden Retriever et particulièrement le Bouvier Bernois, avec une incidence de 25 % dans la race, sont prédisposés au sarcome histiocytaire (Abadie *et al.*, 2009). L'English Springer Spaniel, quant à lui, est très fréquemment atteint de cancers mammaires (Egenvall *et al.*, 2005). Cette particularité présente donc l'avantage d'accélérer le processus de découverte de gènes impliqués dans la prédisposition de ces cancers dans une perspective d'oncologie comparée.

De nombreuses études cliniques et analyses d'histologie comparée montrent la forte homologie qui existe entre les cancers humains et canins, notamment pour les mélanomes, les lymphomes, les leucémies, l'ostéosarcome, les sarcomes des tissus mous, et les carcinomes mammaires, de la prostate, du poumon, de la tête, du cou et de la vessie (Paoloni and Khanna, 2008; Tonomura *et al.*, 2015; Saba *et al.*, 2016).

3. Les mélanomes

3.1. Les principaux types de mélanomes

Le cancer est une cause majeure de décès dans le monde, correspondant à environ 13 % de la mortalité mondiale. Il s'agit d'une maladie complexe aux facteurs multiples, génétiques, liés à l'environnement et au mode de vie et possédant une variabilité inter-individus. Il est évalué aujourd'hui que plus de 2 000 gènes pourraient jouer un rôle dans l'apparition ou le développement d'un cancer (Ilie *et al.*, 2014).

Dans notre étude, nous nous sommes intéressés à un type de cancer, le mélanome. Ce cancer est fréquent en France avec, en 2017, plus de 8 100 nouveaux cas de mélanome cutané touchant les hommes et 7 300 touchant les femmes, soit 3,9 % des cancers, parmi lesquels près de 1 040 décès ont été enregistrés chez l'homme et 750 chez la femme (Institut national du Cancer, 2017). En parallèle, outre-Atlantique, le nombre de nouveaux cas de mélanomes cutanés enregistrés est de plus de 87 000 en 2017, soit environ 5 % de la totalité des cancers aux Etats-Unis, parmi lesquels 9 700 décès ont été recensés. C'est le cancer de la peau le plus agressif, avec une incidence qui augmente rapidement passant d'environ 10/100 000 hommes en 1975 à 30/100 000 hommes en 2013 et d'environ 9/100 000 femmes en 1975 à 15/100 000 femmes en 2013 (Siegel, Miller and Jemal, 2017).

Nous pouvons distinguer deux grands sous-types de mélanomes, d'une part les mélanomes cutanés, souvent dépendants des effets des rayons ultra-violets (UV), et d'autre part, des formes de mélanomes qui se développent à partir de zones peu ou non exposées aux UV.

Les mélanomes cutanés sont fréquents et très étudiés. Les mutations de l'ADN de ces tumeurs sont le plus souvent induites par les rayons UV qui vont affecter les mélanocytes responsables de la production de la mélanine, le pigment de la peau. La tumeur qui se développe se situe classiquement sur la peau, et le mélanome peut alors, en l'absence de diagnostic, s'étendre le long de l'épiderme puis pénétrer les parties plus profondes de la peau. À un stade plus avancé, la tumeur peut entrer en contact avec les vaisseaux lymphatiques et sanguins. Parmi les 3 principaux types de mélanomes cutanés, le plus fréquent est le mélanome avec propagation superficielle (SSM pour 'Superficial Spreading Melanoma'), il représente environ 70 % des mélanomes cutanés. Le mélanome nodulaire,

10 à 15 % de tous les mélanomes, se propage en profondeur plus précocement et est associé à un moins bon pronostic que les autres types de mélanomes. Le mélanome de Dubreuilh, environ 5 % des mélanomes, se développe à partir d'un lentigo pré-existant (petite tâche de la peau, pigmentée et ronde), le plus souvent sur le visage et autres zones exposées au soleil de façon chronique.

En ce qui concerne les mélanomes de zones peu ou non exposées aux UV, il en existe plusieurs types. Le mélanome acrolentigineux se développe typiquement sur les paumes, les plantes, et sous les ongles et représentent moins de 5 % de tous les mélanomes. Il s'agit cependant des mélanomes les plus fréquents chez les populations noires et asiatiques. Le mélanome uvéal (choroïdien, irien et des corps ciliaires) est la tumeur intraoculaire la plus fréquente chez l'homme. C'est une tumeur relativement rare puisque seulement 500 à 600 nouveaux cas sont diagnostiqués chaque année en France. Bien qu'avec les traitements actuels, le contrôle tumoral soit estimé à 95 %, aucune thérapie n'a permis d'améliorer significativement le pronostic vital qui est de 50 % de décès à 15 ans. Les métastases sont hépatiques dans 90 % des cas. Ce tropisme hépatique reste encore mal expliqué et est spécifique du mélanome uvéal. Enfin, le mélanome muqueux se développe à partir des muqueuses de la cavité buccale, nasale, gastro-intestinale, vulvo ou ano-rectale. Les mélanomes muqueux ne sont pas dépendants des UV, peu fréquents mais considérés comme une forme agressive de mélanome. S'il est considéré comme agressif, c'est parce qu'il est souvent découvert à un stade avancé, où les options de traitement sont limitées (Mikkelsen *et al.*, 2016; Hernandez *et al.*, 2018). Il évolue alors généralement en métastase peu de temps après le diagnostic. Par conséquent, les perspectives suivant le diagnostic d'un mélanome muqueux sont sombres.

Au niveau génétique, les principales différences reportées entre les types de mélanome UV et non-UV dépendants concernent les mutations identifiées et leurs incidences. Une proportion significative de mélanomes cutanés présente des mutations récurrentes dans les gènes (i) *BRAF* (environ 50 %), (ii) *NRAS* (environ 20 %) et (iii) *NF1* (environ 17 %) (Hélias-Rodzewicz *et al.*, 2015; Guo, Yao and Jiang, 2016; Hayward *et al.*, 2017; Hernandez *et al.*, 2018). Ces mutations peuvent être associées à l'activation constitutive de la voie de signalisation MAPK ('Mitogen Activated Protein Kinase'). De plus, les mutations de *BRAF* et *NRAS* sont souvent mutuellement exclusives (Hélias-Rodzewicz *et*

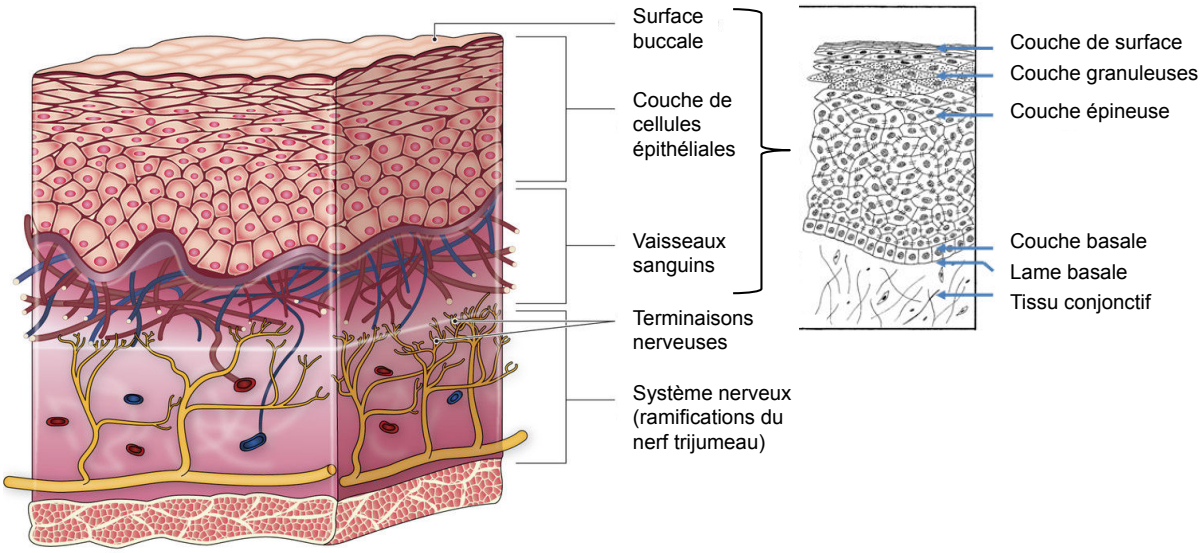

Figure 10 : Schéma de la muqueuse. *A droite correspond une coupe de la muqueuse buccale, extrait de http://monde.ccdmd.qc.ca/ressource/?id=91656. A gauche correspond à un zoom de cette coupe de la muqueuse buccale, composition de l'épithélium, extrait de http://dermatologiebuccale-nice.fr/anatomie-et-histologie-de-la-muqueuse-buccale/histologie-de-la-muqueuse-buccale/.*

al., 2015). Les différentes études des mélanomes muqueux montrent des altérations au niveau des gènes (i) *TP53* (environ 30 %) (Gwosdz *et al.*, 2006), (ii) *PTEN* (environ 48 %), (iii) *p16 / INK4a* (environ 55 %) (Turri-Zanoni *et al.*, 2013) et enfin (iv) *CDKN2A* (entre 30 et 50 %) (Furney *et al.*, 2013; Hsieh *et al.*, 2013; Hayward *et al.*, 2017), et l'absence de mutations de *BRAF* (Hayward *et al.*, 2017).

Les mélanomes muqueux représentent 1,3 % de tous les mélanomes et peuvent se développer dans n'importe quelle membrane muqueuse (Chang, Karnell and Menck, 1998). Nous identifions et dénombrons quatre sous-types de mélanomes muqueux, (i) les mélanomes conjonctivaux (0,5 / million / an), (ii) les mélanomes de la cavité sino-nasale (0,5 / million / an), (iii) les mélanomes ano-rectaux (0,4 / million / an) et (iv) les mélanomes de la cavité buccale (0,2 / million / an). Le mélanome muqueux survient le plus souvent après les 60 ans d'un patient (Mikkelsen *et al.*, 2016). Aux premiers stades, le mélanome muqueux est plat ou légèrement surélevé. Aux stades ultérieurs, la tumeur est élevée, brune à noire et envahit souvent les tissus sous-jacents. L'ulcération est fréquente et a été suggérée comme un facteur pronostic négatif dans les mélanomes sino-nasaux (Mikkelsen *et al.*, 2016). Le traitement chirurgical des mélanomes de la muqueuse est souvent associé à une morbidité sévère, d'où la nécessité d'étudier d'autres traitements.

3.2. Le mélanome muqueux

3.2.a. Les mélanocytes des muqueuses

Les mélanocytes, qui jouent un rôle physiologique important dans la peau, sont également retrouvés dans d'autres épithéliums et notamment au niveau de certaines muqueuses. Les muqueuses désignent les structures épithéliales caractérisées par la sécrétion du mucus, une substance aqueuse épaisse, riche en glycoprotéines, les mucines, et en enzymes antiseptiques comme l'hydrolase. Le mucus permet, entre autres, de faciliter les échanges entre le milieu extérieur et les cellules épithéliales et fournit une protection contre les agents pathogènes. La présence de mélanocytes a été démontrée dans de nombreuses muqueuses et notamment dans les cavités buccales et nasales, dans la face interne des paupières, au niveau des muqueuses sexuelles et anales et même au niveau de l'œsophage.

La muqueuse buccale est un épithélium stratifié dont la structure histologique est relativement similaire à celle de la peau (Figure 10). Au niveau de la face extérieur de la muqueuse, nous retrouvons un épithélium multicouches, essentiellement composé de kératinocytes mais également de mélanocytes et reposant sur une lame basale. L'organisation des couches supra-basales varie en fonction de la localisation anatomique. Au niveau des lèvres, du palais mou et des joues, l'épithélium est non kératinisé. La couche granuleuse ne présente pas ou peu de grains caractéristiques de kératohyaline (protéine structurale qui intervient dans le processus de maturation de l'épiderme) et le processus de dégradation des kératinocytes reste inachevé : les cellules s'aplatissent mais conservent leurs organelles. Dans les régions participant à la mastication : la langue, les gencives et le palais dur, l'épithélium est dit para-kératinisé, il s'agit d'un processus intermédiaire où les cellules présentent des degrés variables de transformations. Le tissu conjonctif sous-jacent appelé chorion est composé de façon similaire au derme de la peau, d'une dense trame de protéines fibreuses. Il possède un rôle essentiellement structural et contient notamment les glandes salivaires.

Dans l'épithélium muqueux, les mélanocytes se localisent au niveau de la couche basale, cependant, alors que leur répartition est irrégulière dans la peau, les mélanocytes des muqueuses sont distribués de façon homogène pour un ratio d'environ un pour 15 kératinocytes. Leur fonction principale de photoprotection, ne s'applique bien sûr pas pour les cavités anatomiques non exposées au rayonnement solaire, leur rôle au sein de ces tissus est donc encore mal compris. D'autant plus que les muqueuses humaines non pathologiques sont en général non pigmentées même si de faibles quantités de mélanosomes (organite synthétisant la mélanine) peuvent être observées dans les kératinocytes. Des hypothèses ont été formulées par extrapolation des connaissances sur les mélanocytes cutanés. Nous pouvons notamment suspecter un rôle possible dans la réponse immunitaire ou dans la détoxification, la mélanine permettant de piéger des substances nocives auxquelles les muqueuses buccales sont fréquemment exposées (Feller *et al.*, 2014).

3.2.b. Le mélanome muqueux chez l'homme

Les mélanomes muqueux sont des tumeurs rares et très agressives, d'étiologie incertaines, touchant les épithéliums muqueux de différentes régions anatomiques. Bien

que le rôle des mélanocytes dans ces tissus non pigmentés soit obscur, leur présence est donc bien établie. Les mélanomes muqueux sont un groupe de tumeurs représentant 1 à 2 % du total des mélanomes. Les sites les plus communément touchés sont les muqueuses de la cavité buccale, du rectum et de l'appareil génital, avec environ 55, 23 et 18 % des cas, respectivement. Peu d'études à grande échelle sont venues renseigner l'épidémiologie de ces cancers. Il semble que tous les phototypes soient touchés de manière égale, même si, conséquence de la faible incidence des mélanomes cutanés dans les populations non caucasiennes, les formes muqueuses représentent une proportion plus importante des mélanomes développés par ces populations. L'âge médian d'apparition se situe dans la septième décade de la vie (Chang, Karnell and Menck, 1998; Marcus *et al.*, 2012). Comme les mélanomes cutanés d'origines épidermiques, ces néoplasmes prennent naissance au niveau de la lame basale à l'interface entre l'épithélium et le tissu conjonctif sous-jacent. Ils présentent en général une phase de croissance horizontale importante formant de larges mélanoses, c'est-à-dire une accumulation anormale de mélanine dans les tissus, avant de produire un nodule évoluant rapidement. À cause de leur localisation anatomique inhabituelle et de l'absence de symptôme précoce, les mélanomes muqueux sont virtuellement tous à un stade très avancé au moment du diagnostic. Les quelques études de suivi de cohorte montrent un taux de suivie à cinq ans dramatiquement bas, toujours inférieur à 25 %. La détection tardive et la richesse du réseau vasculaire et lymphatique des muqueuses, zones d'échanges avec le milieu extérieur, font que 75 % des patients présentent des métastases lymphatiques locales au moment du diagnostic (Patrick, Fenske and Messina, 2007). Ces caractéristiques font du mélanome muqueux la forme la plus grave des proliférations mélanocytaires.

3.2.c. Le mélanome muqueux chez le chien

L'investigation du mélanome muqueux est difficile en raison notamment de la relative rareté de la maladie chez l'homme, et pour lesquelles des études multicentriques et des coopérations entre différentes spécialités médicales sont nécessaires (Mikkelsen *et al.*, 2016; Hernandez *et al.*, 2018). De plus, une étude (Hernandez *et al.*, 2018) a montré que les tumeurs de l'homme comme celles du chien sont généralement résistantes à la chimiothérapie et à la radiothérapie. Le mélanome muqueux est une tumeur relativement

commune chez les chiens (Smedley *et al.*, 2011), avec jusqu'à 100 000 diagnostics par an aux Etats-Unis (Bosenberg, Arnheiter and Kelsh, 2014; Poorman *et al.*, 2015). Le mélanome oral représente la forme la plus fréquente de mélanome chez le chien avec près de 62 % des mélanomes, ce qui en fait un bon modèle pour les mélanomes muqueux humains. L'âge moyen des chiens atteints de néoplasmes mélanocytaires bénins et malins au moment du diagnostic est de 8,1 et 11,6 ans, respectivement. Les chiens affectés de mélanome muqueux ont un temps de survie médian allant de 3 à 18 mois, selon le stade au diagnostic (MacEwen *et al.*, 1986; Smedley *et al.*, 2011). Une étude réalisée par Gillard *et al*., au sein de notre équipe (Gillard *et al.*, 2014), avait permis l'analyse de 2 350 cas de tumeurs mélanocytaires. Il avait été démontré l'existence de prédispositions raciales selon le type de mélanome. Sur l'ensemble des tumeurs mélanocytaires, le Caniche est prédisposé aux mélanomes de la cavité buccale avec près de 77 % des cas de mélanomes. Les Bergers de Beauce, les Rottweillers et les Schnauzers, quant à eux, sont prédisposés aux mélanomes cutanés et digités avec 80 %, 69 % et 63 % des cas de mélanomes, respectivement. D'autres races, telles que le Labrador, le Golden Retriever ou le Boxer, ne présentent pas particulièrement de spécificité de site. Une autre étude met en évidence d'autres races comme le Terrier Écossais, le Teckel et le Cocker Anglais (Bergman, Kent and Farese, 2013), tout en indiquant également le Caniche et le Golden Retriever. Parmi ces races de chiens prédisposées, il a été mis en évidence une surreprésentation des tumeurs chez les races à poils noirs, telles que les Schnauzers, les Rottweilers ou les Bergers de Beauce (Gillard *et al.*, 2014). Au contraire, les races à poils pâles, blancs et mêmes celles sans poils, apparaissaient sous-représentées dans l'ensemble de données de cette étude et ne semblent pas prédisposées à développer un mélanome muqueux ou cutané. L'étude met également en avant la sévérité des mélanomes muqueux par rapport aux mélanomes cutanés chez le chien.

Dans le travail de Gillard *et al.*, de 2014 (Gillard *et al.*, 2014), une analyse des mutations somatiques canines, dans le cas de mélanomes oraux, avait été réalisée en criblant six gènes qui contiennent des mutations somatiques récurrentes dans différents sous-types de mélanomes humains, *BRAF*, *NRAS*, *PTEN*, *KIT*, *GNAQ* et *CDK4*. Sur un ensemble de 35 mélanomes canins (27 oraux et 8 cutanés), aucune altération n'avait été retrouvée dans les gènes *BRAF*, *GNAQ* et *CDK4*, ni dans le gène *KIT*, concordant avec l'étude de

Murakami *et al.*, (Murakami *et al.*, 2011). L'absence d'altération du gène *BRAF* dans le mélanome muqueux et cutané est en accord avec le fait que, d'une part le mélanome muqueux humain ne présente pas de mutation de *BRAF* et, d'autre part, que le mélanome cutané canin se comporte différemment du mélanome cutané humain. Ces résultats ont été confortés dans l'étude de Mochizuki *et al.*, de 2015 (Mochizuki *et al.*, 2015), ou seuls 3 de 54 échantillons (5 %) de mélanomes cutanés canins présentent une mutation correspondant à la mutation la plus commune de *BRAF* dans les cancers humains, V600E, présente dans plus de 60 % chez l'homme. Cette dernière observation peut s'expliquer par le fait que la mutation *BRAF* est observée en présence d'UV et que les chiens présentent des mélanomes cutanés dermiques sur des sites protégés de l'exposition au soleil grâce à leurs poils.

Dans le mélanome muqueux canin, les deux gènes *NRAS* et *PTEN*, ont été identifiés comme porteurs de mutations somatiques, avec un effet dommageable sur la protéine : *NRAS*, aux positions Q61 et G12, et *PTEN* à la position G251. Les mutations *NRAS* correspondent aux mutations humaines activatrices des codons Q61 et G12 portant la plupart des mutations *NRAS* des mélanomes humains. Ces résultats confirment la pertinence de l'étude des mélanomes muqueux canins comme modèles naturels pour les mélanomes muqueux homologues humains.

L'Organisation Mondiale de la Santé a défini un schéma basé sur la taille de la tumeur pour les chiens atteints de mélanome muqueux, stade I = < Tumeur de 2 cm de diamètre, stade II = Tumeur de 2 à < 4 cm de diamètre, stade III = > Tumeur de 4 cm et / ou métastase ganglionnaire et stade IV = métastase à distance. Les durées médianes de survie des chiens atteints d'un mélanome muqueux traité par chirurgie sont respectivement d'environ 17 à 18, 5 à 6 et 3 mois avec une maladie de stade I, II et III (MacEwen *et al.*, 1986). Des rapports plus récents (Freeman *et al.*, 2003; Proulx *et al.*, 2003) suggèrent que le mélanome muqueux de stade I traité avec des thérapies standardisées comprenant chirurgie, radiothérapie et / ou chimiothérapie a une durée médiane de survie d'environ 12 à 14 mois, la plupart des chiens mourant de métastases à distance et non de récidive locale. D'autres chercheurs (Théon, Rodriguez and Madewell, 1997) ont reporté que les chiens atteints d'un mélanome muqueux de stade I présentaient des durées médianes de survie sans progression de 19 mois.

Dans l'industrie pharmaceutique, les chiens sont donc de plus en plus inclus dans les essais précliniques, reflétant ainsi les essais cliniques humains (Paoloni and Khanna, 2008;

Smedley *et al.*, 2011). Les chiens constituent un avantage pour les informations pharmacocinétiques et pharmacodynamiques (Hernandez *et al.*, 2018). La taille des organes est comparable entre les races de grande taille et l'humain, ce qui rend le chien très utile dans les essais cliniques car les procédures chirurgicales sont relativement similaires entre l'homme et le chien. Actuellement, le chien est principalement utilisé pour des essais de médicaments anti-cancéreux dans le but d'établir les profils de sécurité nécessaires à l'utilisation des molécules dans les essais de phase I chez l'homme. Cependant, depuis une dizaine d'années, le chien atteint de cancer est progressivement inclus dans l'évaluation de l'efficacité de nouveaux traitements (Hansen and Khanna, 2004). Car, en plus de présenter des tumeurs comparables aux tumeurs humaines sur les plans clinique, histologique et génétique, le chien présente également la même réponse aux traitements (Paoloni and Khanna, 2008). D'ailleurs, les protocoles et traitements déjà utilisés pour les cancers du chien sont bien souvent calqués sur ceux utilisés chez l'homme. Enfin, il existe des techniques aussi élaborées que celles développées pour les soins des cancers humains et les structures capables de prendre en charge les chiens atteints de cancers se sont grandement développées, aux Etats-Unis comme en Europe.

Résultats

Les résultats sont présentés sous la forme de 3 parties. Une première partie présente le travail de réannotation du génome canin dans le cadre d'une étude réalisée en 2017 (Wucher *et al.*, 2017), qui constitue la base des données de l'annotation des gènes canins sur lesquelles repose le travail de thèse. La seconde partie décrit la caractérisation de l'expression des ARNs longs non-codants canins (lncRNAs), leur comparaison avec les gènes codant pour des protéines (mRNAs), ainsi que leur conservation avec l'homme et l'exploration de leurs potentiels rôles fonctionnels. Ces résultats font l'objet d'un article publié dans le journal *Scientific Reports* (Le Béguec *et al.*, 2018). La dernière partie présente les résultats préliminaires qui identifient et évaluent l'impact des lncRNAs dans la biologie d'un cancer, le mélanome oral canin comme modèle du mélanome muqueux humain.

I. Identification du répertoire de lncRNAs

1. FEELnc, un programme d'annotation des lncRNAs

Début 2017, une extension de l'annotation du génome canin ainsi qu'un outil bioinformatique qui annote précisément les lncRNAs, ont été réalisés par l'équipe et publiés dans l'article de Valentin Wucher (Wucher *et al.*, 2017). Ce programme nommé FEELnc, pour 'Flexible Extraction of LncRNAs', permet (i) de filtrer les modèles de transcrits assemblés par RNA-seq (module FEELnc$_{_filter}$), (ii) de discriminer les transcrits non-codants (lncRNAs) des transcrits codant pour des protéines (mRNAs) via le calcul d'un potentiel codant (module FEELnc$_{_codpot}$), et (iii) de classer ces nouveaux transcrits selon leurs positions génomiques et leurs orientations de transcription (module FEELnc$_{_classifier}$). FEELnc est un programme qui appartient à la catégorie des outils sans alignement ('alignment-free') et qui annote précisément les transcrits en s'appuyant sur un modèle de forêt aléatoire ('Random Forest') (Breiman, 2001) entraîné sur des caractéristiques inhérentes aux séquences telles que la couverture des cadres ouverts de lecture (ORFs), la taille des transcrits et les fréquences de *k*-mer.

À partir de l'ensemble des modèles de gènes définis par les outils de reconstruction de transcrits tel que Cufflinks (Trapnell *et al.*, 2012), ce sont 3 variables (ou prédicteurs) qui sont caractérisés dans un premier temps par FEELnc et plus spécifiquement par le module

FEELnc$_{_codpot}$. Le premier prédicteur est la longueur des transcrits car des études précédentes ont montré que les séquences de lncRNAs étaient significativement plus petites que les mRNAs (Derrien, Johnson, *et al.*, 2012). Le second prédicteur est la fréquence des oligonucléotides ou *k*-mer, où *k* désigne la taille de l'oligonucléotide, utilisée comme signature des séquences codantes versus non-codantes. Certains outils utilisent déjà des fréquences de *k*-mers pour la discrimination des lncRNAs par rapport aux mRNAs mais sont souvent limités à des *k*-mers court (généralement $k \leq 6$) (Wang *et al.*, 2013), alors que les *k*-mers plus longs peuvent aider à résoudre les ambiguïtés en tenant compte des répétitions spécifiques aux lncRNAs. FEELnc assure l'analyse très rapide de multiples *k*-mer ($k = 1$ à 12), et examine les cadres de lectures possibles des ORFs. Le 3ème prédicteur est la couverture des ORFs ('Open Reading Frames' ou cadre ouvert de lectures) (Brent, 2007), car plus l'ORF est grand, plus grande est la probabilité que le transcrit soit codant pour une protéine. Si des jeux d'apprentissage codant (mRNA) et non-codant (lncRNA) sont disponibles pour l'organisme d'intérêt, les valeurs des 3 types de prédicteurs sont déterminées pour l'ensemble des séquences de transcrits et sont ensuite incorporées dans une méthode d'apprentissage automatique ('machine learning') utilisant le principe des forêts aléatoires ('Random Forests') (Breiman, 2001) pour classer les lncRNAs et les mRNAs. Il est à noter que si l'espèce d'intérêt ne dispose pas de jeu d'apprentissage non-codant (ce qui est souvent le cas des espèces 'non-modèle'), FEELnc peut en dériver un par 'pseudogénéization' du jeu de données codant c'est-à-dire en réalisant une distribution aléatoire de la fréquence des *k*-mers des transcrits mRNAs (option '`--mode=shuffle`' du programme). L'étape de Random Forest, qui peut se définir comme un ensemble d'arbres de décision, calcule un score de potentiel codant (ou CPS pour 'Coding Potential Score') pour chaque séquence des jeux d'entraînement. Ce score correspond à la proportion d'arbres parmi tous les arbres *i.e* de la forêt (généralement 500), qui va 'voter' pour qu'une séquence du jeu d'apprentissage soit codante ou non-codante. L'algorithme Random Forest présente un triple avantage car (i) il permet de gérer des données d'apprentissage non équilibrées ce qui est généralement le cas des annotations codant versus non-codant où une majorité de mRNAs sont connus et bien caractérisés tandis que, en général, encore peu de lncRNAs sont connus et bien annotés chez certains organismes, (ii) l'algorithme est facilement optimisable et (iii) il peut analyser des matrices creuses, c'est-à-dire incluant des données de valeur nulle. À la fin de l'étape

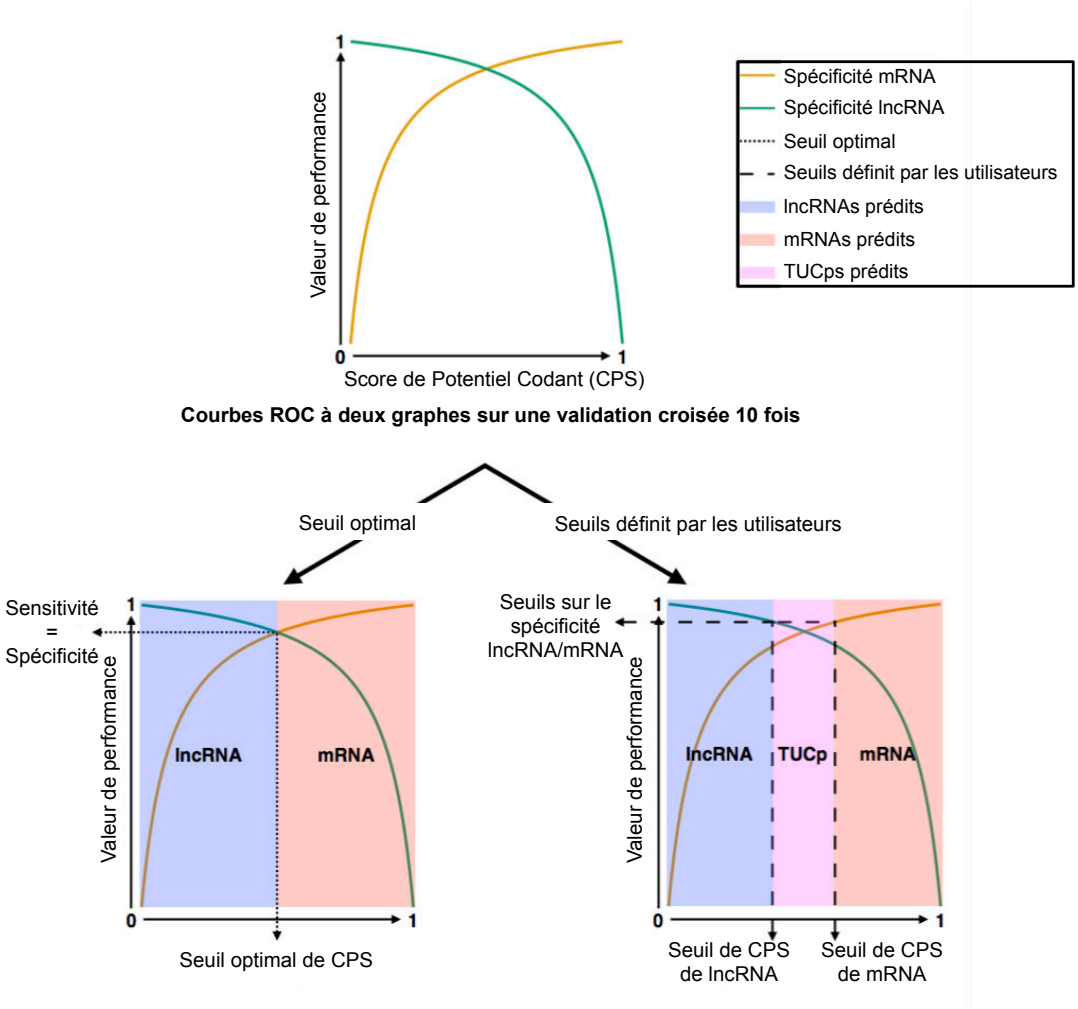

Figure 11 : Description de FEELnc$_{_codpot}$ et FEELnc$_{_classifier}$. *Courbes ROC à deux courbes pour la détection automatique du seuil CPS optimisé et du seuil de spécification de l'utilisateur, ce dernier définissant deux ensembles de lncRNAs et mRNAs et une classe de transcrits avec des biotypes ambigus dits TUCp ('Transcriptions du potentiel de codage inconnu'). Traduit de Wucher et al., 2017.*

d'entraînement, chacune des séquences des jeux d'apprentissage a donc un score CPS et la dernière étape du module FEELnc$_{_codpot}$ consiste à déterminer le seuil de CPS qui permet de mieux discriminer les séquences codantes des séquences non-codantes. Pour ce faire, l'outil construit automatiquement une courbe ROC ('Receiver Operating Characteristic') où les valeurs de performances de l'étape d'entraînement en termes de Sensibilité (Sn) et Spécificité (Sp) sont calculées pour l'ensemble des seuils de CPS (Figure 11). Le seuil de CPS qui maximise à la fois la sensibilité et la spécificité, en d'autres termes quand Sn == Sp, est ainsi choisi et sera utilisé comme seuil pour différentier les lncRNAs des mRNAs sur les données testées. Nous pouvons enfin mentionner la possibilité offerte par FEELnc de fixer les seuils de spécificité (lncRNAs et mRNAs) afin de définir des jeux de données mRNAs et lncRNAs plus robustes et conservatifs. Cette dernière option entraînera la définition d'une classe intermédiaire entre lncRNA et mRNA, appelée TUCp pour 'Transcript of Unknown Coding potential', pour laquelle le potentiel codant calculé est dans l'intervalle entre celui des lncRNAs et mRNAs (Figure 11).

2. FEELnc et son application au transcriptome canin

Afin d'étendre l'annotation du génome canin, le programme FEELnc a été utilisé sur un ensemble de 20 RNA-seq comprenant des échantillons de tissus et systèmes anatomiques canins collectés par l'équipe et d'autres collaborateurs au sein du consortium européen LUPA (Lequarré *et al.*, 2011) et séquencés par les collaborateurs du BROAD Institute of MIT (Cambridge, USA) de l'équipe du Dr. K.Lindblad-toh. Les 20 échantillons comprennent 16 tissus canins distincts (Wucher *et al.*, 2017), inclus dans l'analyse détaillée ci-dessous. La confection des banques et le séquençage de fragment de 100 nucléotides de type 'paired-end' ont été réalisés au BROAD Institute (voir Matériels et Méthodes). L'analyse bioinformatique traitée au laboratoire, a considéré environ 1,3 milliard de reads qui ont été alignées sur l'assemblage CanFam3 à l'aide du programme d'alignement STAR (Dobin *et al.*, 2013) dédié au RNA-seq. L'outil Cufflinks (Trapnell *et al.*, 2010) a été ensuite utilisé pour reconstruire les modèles de transcrits pour chacun des 20 RNA-seq séparément guidés par une annotation de référence consensuelle donnée par la combinaison des annotations du BROAD (Hoeppner *et al.*, 2014) et d'EnsEMBL (version v83) (Yates *et al.*, 2016), appelée CanFam3.1. Puis c'est l'outil Cuffmerge (Trapnell *et al.*, 2010), qui a été utilisé pour réunir ou

‘merger’ l’ensemble des annotations des 20 jeux de données en un seul fichier. Le module FEELnc$_{_codpot}$ du programme FEELnc a été utilisé pour calculer le potentiel codant des nouveaux transcrits assemblés et ainsi différencier les biotypes codant et non-codant en utilisant le mode ‘shuffle’ de FEELnc. Enfin, les nouveaux transcrits lncRNAs/mRNAs ont été classifiés selon la nomenclature officielle du consortium humain GENCODE (Harrow *et al.*, 2012) en classe intergénique/génique grâce au module FEELnc$_{_classifier}$.

La réannotation du génome canin a mis en évidence un total de 36 327 gènes et 189 114 transcrits de biotypes différents composés de mRNAs, lncRNAs, miRNAs, et pseudogènes, dont 10 374 et 58 640 nouveaux transcrits de lncRNAs et mRNAs. La nouvelle annotation est nommée CanFam3.1-plus et présente 58,6 % de transcrits mRNAs supplémentaires et un doublement du nombre de transcrits de lncRNAs (Cf. Table 1, partie Introduction p23). Ce travail a montré que les nouveaux transcrits de lncRNAs et de mRNAs étaient plus longs et comportaient plus d'exons comparés aux transcripts de la précédente annotation CanFam3.1 ; suggérant la reconstruction plus complète des modèles de gènes. Le nombre d'isoformes par gène a également augmenté allant à ~ 2,2 et ~ 7,2 pour les lncRNAs et les mRNAs, respectivement. Le nombre de loci et de transcrits de lncRNAs détectés par FEELnc étend significativement l'annotation du génome canin fournissant une ressource qui aidera à investiguer les relations entre le génotype et le phénotype (Andersson *et al.*, 2015).

3. Intersection de l’annotation Canfam3.1-plus et des promoteurs et enhancers

Au cours de mon stage de Master 2 réalisé à l’EBI (‘European Bioinformatic Institute’) à Hinxton, j’ai étudié les régions régulatrices de la transcription et plus particulièrement les enhancers et les promoteurs chez 5 mammifères dont le chien (Villar *et al.*, 2015). Par la suite, j’ai pu récupérer les coordonnées génomiques des promoteurs et enhancers canins annotés par Chip-Seq (H3K4me3 et H3K27ac pour promoteurs et enhancers, respectivement) dans le tissue du foie (Villar *et al.*, 2015). Cette étude rapporte que l’évolution rapide des enhancers est une caractéristique universelle des génomes de mammifères et révèlent qu’au contraire les promoteurs sont partiellement ou entièrement conservés dans les 20 espèces étudiées. Cette étude m’a permis de réaliser un contrôle qualité des lncRNAs identifiés par l’outil FEELnc, en analysant l’intersection entre les coordonnées des lncRNAs de l’annotation de CanFam3.1-plus et les promoteurs et

Organes	Race	Nb reads	Données d'origine	Système anatomiques
Bulbe olfactif	Grand Bouvier Suisse	45 799 491	(Wucher et al. 2017)	Système nerveux
Cerveau	Beagle	39 268 738	(Hoeppner et al. 2014)	Système nerveux
Cervelet	Grand Couvier Suisse	44 902 865	(Wucher et al. 2017)	Système nerveux
Cœur	Beagle	24 930 702	(Hoeppner et al. 2014)	S. musculo-squelettiq.
Cortex	Berger Belge	41 319 413	(Wucher et al. 2017)	Système nerveux
Foie	Beagle	25 010 943	(Hoeppner et al. 2014)	'Unmapped'
Follicule pileux	Labrador	45 694 722	(Wucher et al. 2017)	Système tégumentaire
Glandes surrénales	Bouvier Bernois	54 776 586	(Wucher et al. 2017)	'Unmapped'
Glandes mammaires	Grand Bouvier Suisse	44 349 725	(Wucher et al. 2017)	Système tégumentaire
Intestin_côlon	Bouvier Bernois	52 310 396	(Wucher et al. 2017)	Système digestif
Jéjunum	Labrador	50 569 866	(Wucher et al. 2017)	Système digestif
Kératinocyte	Beagle	54 482 221	(Hoeppner et al. 2014)	Système tégumentaire
Moelle épinière	Grand Bouvier Suisse	46 844 306	(Wucher et al. 2017)	Nervous system
Muqueux	Labrador	44 195 944	(Wucher et al. 2017)	Système tégumentaire
Muscle	Beagle	24 808 375	(Hoeppner et al. 2014)	S. musculo-squelettiq.
Pancréas	Berger Belge	47 171 936	(Wucher et al. 2017)	'Unmapped'
Peau	Beagle	52 275 710	(Wucher et al. 2017)	Système tégumentaire
Planum nasal	Labrador	69 193 538	(Wucher et al. 2017)	Système tégumentaire
Poumon	Beagle	22 278 693	(Hoeppner et al. 2014)	'Unmapped'
Rate	Berger Belge	49 604 583	(Wucher et al. 2017)	'Unmapped'
Rein	Beagle	27 691 951	(Hoeppner et al. 2014)	'Unmapped'
Rétine	Border Collie	50 480 134	(Wucher et al. 2017)	'Unmapped'
Sang	Beagle	25 294 750	(Hoeppner et al. 2014)	'Unmapped'
Thymus	Lévrier Persan	51 079 197	(Wucher et al. 2017)	'Unmapped'
Tissu ovarien	Beagle	69 193 538	(Hoeppner et al. 2014)	'Unmapped'
Tissu testiculaire	Beagle	26 743 805	(Hoeppner et al. 2014)	Tissu testiculaire

***Table 3 : Description de chaque type de tissu**, de race de chien, du nombre total de lectures, de référence de publication et de systèmes anatomiques.*

enhancers extraits de Villar *et al*., de 2015 (Villar *et al.*, 2015). Afin de tester la significativité des intersections (contrôle négatif), des coordonnées génomiques de promoteurs ont été simulés ('marques aléatoires') en relocalisant des régions de 5 kb sur le génome canin (voir Matériels et Méthodes). Les marques de promoteurs H3K4me3 chevauchent à 17,5 % les promoteurs des lncRNAs de l'annotation contre ~ 5 % de chevauchement avec les 'marques aléatoires' (test de Wilcoxon, p-value < $2{,}2^{e-16}$). Ces résultats confirment une co-localisation des marques de promoteurs et des promoteurs lncRNAs de l'annotation CanFam3.1-plus. Concernant les lncRNAs chevauchant une marque enhancers, nous observons 2 866 lncRNAs (27,4 %) qui chevauchent au moins une marque H3K27ac, comparé à ~ 15 % en analysant les marques relocalisées aléatoirement (test de Wilcoxon, p-value < $2{,}2^{e-16}$). Sur ces 27,4 %, 1 356 lncRNAs ne chevauchent pas de promoteur, ce qui suggère que ces lncRNAs pourraient avoir un rôle d'enhancer potentiel (eRNA) (Melo *et al.*, 2013).

II. Caractérisation des lncRNAs canins

1. Caractérisation des profils d'expression des lncRNA canins

L'analyse de l'expression de l'ensemble des lncRNAs et des gènes codant pour des protéines (mRNAs) a été réalisée à l'aide du pipeline bioinformatique décrit dans Djebali *et al*., (Djebali *et al.*, 2017), qui réalise l'alignement sur le génome de référence et le comptage des reads sur l'annotation de référence (voir Matériels et Méthode). À partir des données d'expression par tissu, nous avons compilé l'ensemble des niveaux d'expression des lncRNAs et mRNAs en une matrice globale composée des 32 254 gènes codants et des ARN longs non-codants. Pour cette analyse, nous différencions les lncRNAs en deux catégories, les ARN intergéniques longs non-codants (*lincRNAs*, n = 5 651), et les lncRNAs chevauchant en antisens des gènes codant pour des protéines (*antisens*, n = 4 793) en se basant sur la classification réalisée par le module l'outil FEELnc$_{_classifier}$ développé dans l'équipe (Wucher *et al.*, 2017). L'analyse de l'expression des lncRNAs est réalisée à partir de 26 tissus ou systèmes anatomiques canins distincts (Table 3) qui sont représentatifs des principales grandes fonctions de l'organisme (voir Matériels et Méthodes). Les profils d'expression obtenus révèlent que près de 75 % (7 764 / 10 444) des lncRNAs sont exprimés dans au moins un tissu avec une valeur d'expression exprimé en TPM > 1 (Transcripts Per Million) (voir Matériels et Méthodes), unité qui normalise la taille du gène et la profondeur du

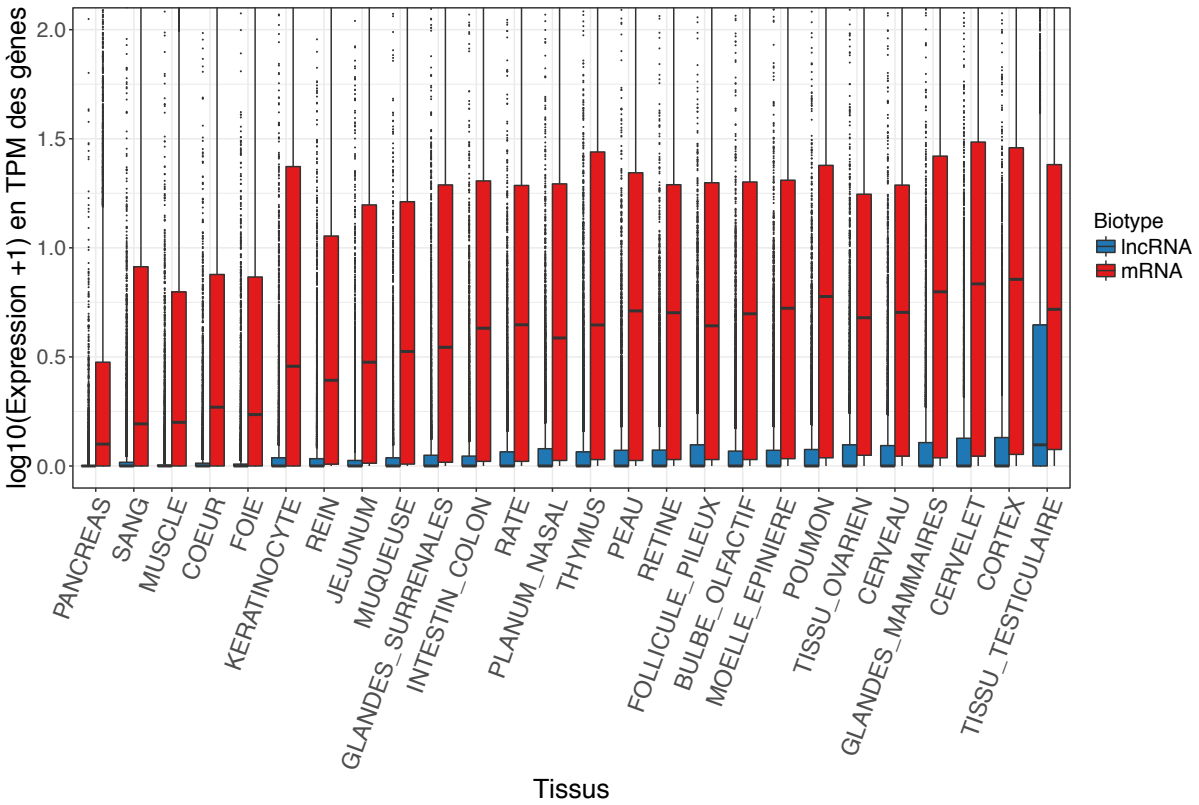

Figure 12 : Analyse comparative des niveaux de transcription entre les mRNAs (rouge) et les lncRNAs (bleu) *en log (TPM +1) dans les 26 tissus canins.*

Organes	lncRNA, moyenne (TPM)	mRNA, moyenne (TPM)
Bulbe olfactif	0,72	45,20
Cerveau	0,80	45,20
Cervelet	1,22	44,80
Cœur	0,39	45,50
Cortex	1,05	44,97
Foie	0,65	45,24
Follicule pileux	1,60	44,71
Glandes surrénales	1,15	44,86
Glandes mammaires	1,43	44,68
Intestin_côlon	0,78	45,10
Jéjunum	0,68	45,03
Kératinocyte	0,73	45,12
Moelle épinière	0,68	45,25
Muqueuse	0,58	45,33
Muscle	0,30	45,60
Pancréas	0,11	45,74
Peau	0,96	44,97
Planum nasal	0,96	44,86
Poumons	0,73	44,98
Rate	0,68	44,94
Rein	0,60	45,30
Rétine	0,69	45,04
Sang	0,43	45,30
Thymus	1,20	44,68
Tissu ovarien	1,29	44,77
Tissu testiculaire	7,02	41,91

Table 4 : TPM_{moyen} pour chaque tissu pour les lncRNAs et les mRNAs.

séquençage. Lorsque nous réalisons cette analyse d'expression pour les mRNAs, 18 730 / 21 810 sont exprimés, soit 86 % exprimés dans au moins un tissu avec une valeur de TPM > 1, un seuil empirique utilisé pour définir un niveau d'expression significatif (Li and Dewey, 2011). Nous avons en parallèle déterminé que 9 542 lncRNAs et 20 434 mRNAs sont exprimés avec une valeur d'expression supérieure à 0,1 TPM, soit 91,4 % et 93,7 % respectivement ; indiquant une bonne représentativité d'expression de la collection de 26 tissus et systèmes anatomiques utilisés dans notre étude.

1.1. Analyse comparée des niveaux d'expression des lncRNAs et des mRNAs

Au delà du grand nombre de lncRNAs (75 %) exprimés (TPM > 1) dans notre étude, nous avons comparé les profils d'expression des lncRNAs et des mRNAs par tissu. Nous observons, comme chez l'homme (Derrien, Johnson, *et al.*, 2012), la souris (Pervouchine *et al.*, 2015) ou le poulet (Muret *et al.*, 2017), que les lncRNAs sont près de 20 fois moins exprimés que les mRNAs (test de Wilcoxon, p-value < $2{,}2^{e-16}$, Figure 12) pour les lncRNAs. Nous notons une exception pour le tissu testiculaire, où, comme il a déjà été démontré chez l'homme ou la souris (Derrien, Johnson, *et al.*, 2012; Soumillon *et al.*, 2013), les lncRNAs sont plus fortement exprimés. Dans notre étude, ils ne sont que six fois moins exprimés que les mRNAs. La forte expression de transcrits dans le tissu testiculaire est en partie liée aux modifications d'histone qui favorisent l'ouverture et l'accessibilité de la chromatine (Soumillon *et al.*, 2013) facilitant l'expression des gènes. La moyenne d'expression des lncRNAs par tissu présente une certaine homogénéité allant de 0,11 TPM pour le pancréas à 1,60 TPM pour le tissu du follicule pileux, avec une moyenne de 1,05 TPM, à l'exception du tissu testiculaire (7 TPM), contrairement aux mRNAs où, comme chez l'homme, la moyenne d'expression varie peu, de 41,9 TPM pour le tissu testiculaire à 45,7 TPM pour le pancréas (Table 4).

Nous avons ensuite déterminé la différence d'expression des deux catégories de lncRNAs, les lincRNAs et les antisens. Les antisens sont plus fortement exprimés que les lincRNAs (Figure 13 – 1,20 TPM versus 0,90 TPM pour les antisens et les lincRNAs respectivement, test de Wilcoxon, p-value < $2{,}2^{e-16}$), avec des valeurs pour les antisens comprises entre 0,15 TPM pour le pancréas et 2,0 TPM pour le tissu ovarien, et pour les

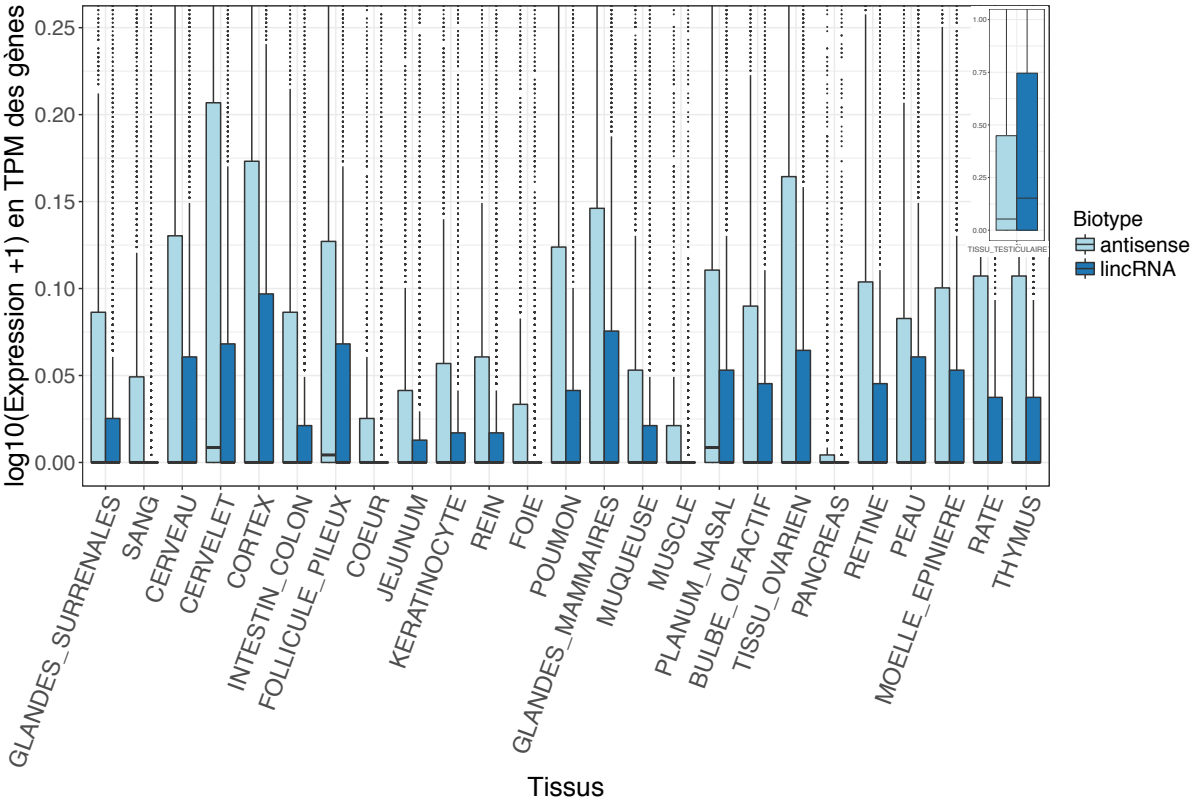

Figure 13 : Analyse comparative des niveaux de transcription entre les antisense (bleu clair) et les lincRNAs (bleu foncé) *en log (TPM +1) dans les 26 tissus canins.*

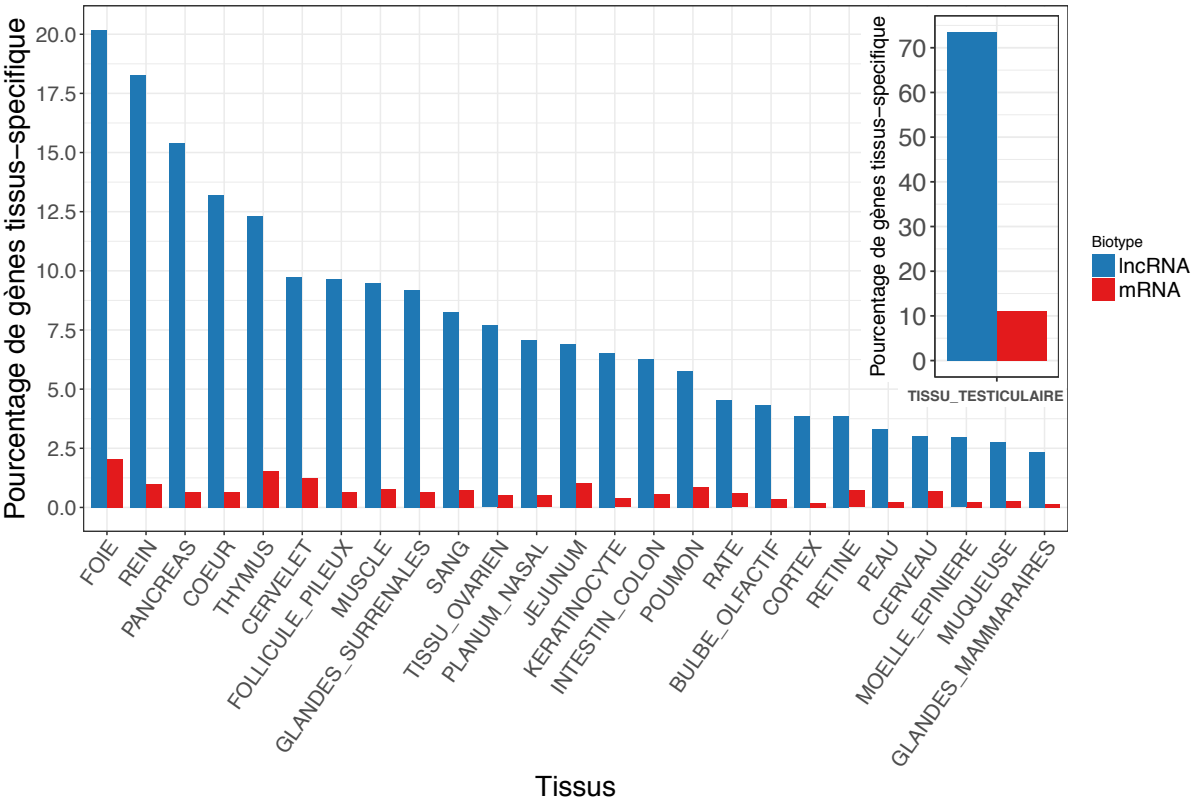

Figure 14 : Proportion des lncRNAs tissus-spécifique (bleu) et des mRNAs tissus-spécifique (rouge) *pour chaque tissu.*

lincRNAs des valeurs comprises entre 0,08 TPM pour le pancréas et 1,72 TPM pour le follicule pileux. Comme précédemment, le tissu testiculaire fait exception, où ce sont les lincRNAs les plus exprimés, 8,0 TPM et 5,5 TPM pour les lincRNAs et les antisens respectivement (test de Wilcoxon, p-value < $2{,}2^{e-16}$).

1.2. Analyse de la tissu-spécificité des lncRNAs

Les lncRNAs ont été caractérisés dans les autres espèces par des profils d'expression qui indiquent qu'ils sont exprimés généralement dans un seul tissu ou à un petit groupe de tissus. L'analyse de la répartition de l'expression des lncRNAs au sein des tissus se détermine à l'aide d'une mesure adaptée, un score de spécificité, dont l'objectif est de quantifier la spécificité d'expression tissulaire d'un gène lncRNA ou mRNA. Dans cette étude, nous avons utilisé le score de spécificité appelé 'tau' dénoté $\boldsymbol{\tau}$ (Yanai *et al.*, 2005). Ce score est compris entre 0 et 1, où 1 correspond à l'expression d'un gène dans un seul tissu et zéro correspondrait à une expression ubiquitaire dans les tissus. Nous avons calculé $\boldsymbol{\tau}$ pour l'ensemble des gènes et avons utilisé le seuil de $\boldsymbol{\tau} > 0{,}95$ pour définir un gène tissu-spécifique. Ce seuil met en évidence des gènes étant en moyenne quatre fois plus exprimés dans le tissu dit spécifique que dans le tissu ayant une valeur d'expression un rang en dessous (voir Matériels et Méthodes). En utilisant ce seuil, nous identifions près de 44 % de lncRNAs exprimés de manière spécifique à un tissu ou à un groupe limité de tissus ($\boldsymbol{\tau} \geq 0{,}95$), que nous définissons comme des lncRNAs tissu-spécifiques (n = 4 599) c'est-à-dire dont l'expression et la fonction sont préférentiellement assignées à un ou quelques types tissulaires ou cellulaires. En comparaison, seuls 17 % des mRNAs (n = 3 635) sont tissu-spécifiques. Cette comparaison reflète comme chez d'autres espèces, que les lncRNAs canins sont fortement tissu-spécifiques et significativement plus que les mRNAs (Kornienko *et al.*, 2016). Ces résultats suggèrent ainsi de potentielles fonctions spécialisées des lncRNAs.

Nous avons ensuite déterminé la proportion des lncRNAs tissu-spécifiques par tissu puis comparé ces résultats aux valeurs obtenues pour les mRNAs. Pour chaque tissu, la proportion de lncRNAs exprimés spécifiquement varie de 2,3 à 20,2 % alors que pour les mRNAs, la proportion de gènes tissu-spécifiques varie de 0,16 à 2,0 % (Figure 14). Le tissu testiculaire présente un écart encore plus important avec plus de 70 % des lncRNAs

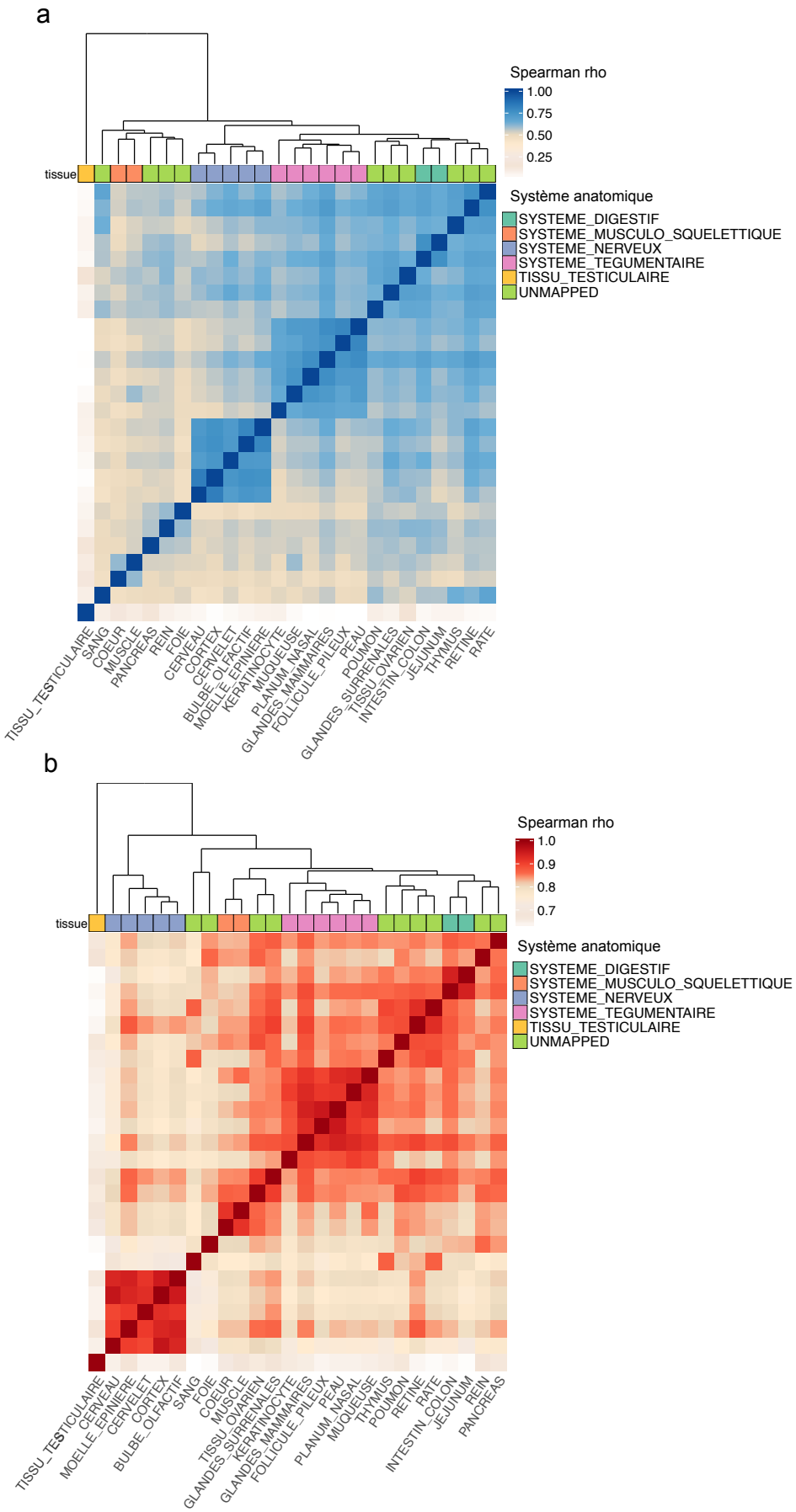

Figure 15 : Clustering hiérarchiques ascendants des 26 tissus *basé sur une corrélation de Spearman mesurée à partir des données d'expression des lncRNAs ((a) bleu) et des mRNAs ((b) rouge).*

exprimés dans le tissu testiculaire (TPM > 1) de manière spécifique ($\boldsymbol{\tau} \geq 0{,}95$, $TPM_{_moyen} = 21{,}20$), correspondant à 3 001 / 4 599 lncRNAs, alors que cette proportion n'est que de 10 % pour les mRNAs. À l'exception du tissu testiculaire, nous avons identifié une moyenne de 63 lncRNAs tissu-spécifiques pour les 25 tissus. L'analyse des échantillons correspondant aux différents systèmes anatomiques nerveux (voir Matériels et Méthode), incluant le cervelet, le cortex et le bulbe olfactif, permet d'identifier un ensemble de 260 lncRNAs présentant des profils d'expression relativement spécifiques du système nerveux. Ces derniers sont répartis en 156 lncRNAs spécifiquement exprimés dans le cervelet, 60 lncRNAs dans le cortex et 44 lncRNAs dans le bulbe olfactif. En moyenne, ces 260 lncRNAs ont un fort niveau d'expression ($TPM_{_moyen} = 3{,}30$) en comparaison de la valeur moyenne des lncRNAs dans les 23 autres tissus ($TPM_{_moyen} = 0{,}12$). À titre d'exemple, nous avons caractérisé un lincRNA spécifiquement exprimé dans le cervelet ($\boldsymbol{\tau} = 0{,}96$) qui est annoté chez l'homme comme étant le lncRNA *CASC18 ('Cancer Susceptibility Candidate gene 18')*. Dans la littérature, *CASC18* est décrit comme ayant un rôle dans la différenciation de cellules neurales (Mehravar *et al.*, 2017), ce qui apparaît en accord avec son patron d'expression restreint aux tissus nerveux.

1.3. Analyse des données d'expression par clustering ascendant hiérarchique

L'analyse des données d'expression des lncRNAs par les méthodes utilisant des scores de spécificité permet de définir que leurs patrons d'expression sont restreints à un ou à un petit de nombre de tissus, ce qui suggère une fonction relativement spécialisée à un ou un faible nombre de tissus. La détermination des relations entre les tissus basée sur l'analyse de l'expression des gènes a été bien formalisée par des méthodes de clustering telle que la méthode décrite dans Breschi *et al*., (Breschi *et al.*, 2016). Nous avons réalisé une analyse de clustering ascendant hiérarchique à partir de l'ensemble des profils d'expression des lncRNAs des 26 tissus. Nous avons calculé une matrice de distance, grâce à une corrélation de Spearman par paires entre les 26 tissus, à partir de laquelle nous avons produit une carte de chaleur ('heatmap') avec la fonction `geom_tile` de la librairie ggplot2 et un clustering des échantillons par la fonction R `hclust` (distance euclidienne et méthode d'agrégation dite 'complete', voir Matériels et Méthodes ; Figure 15a.).

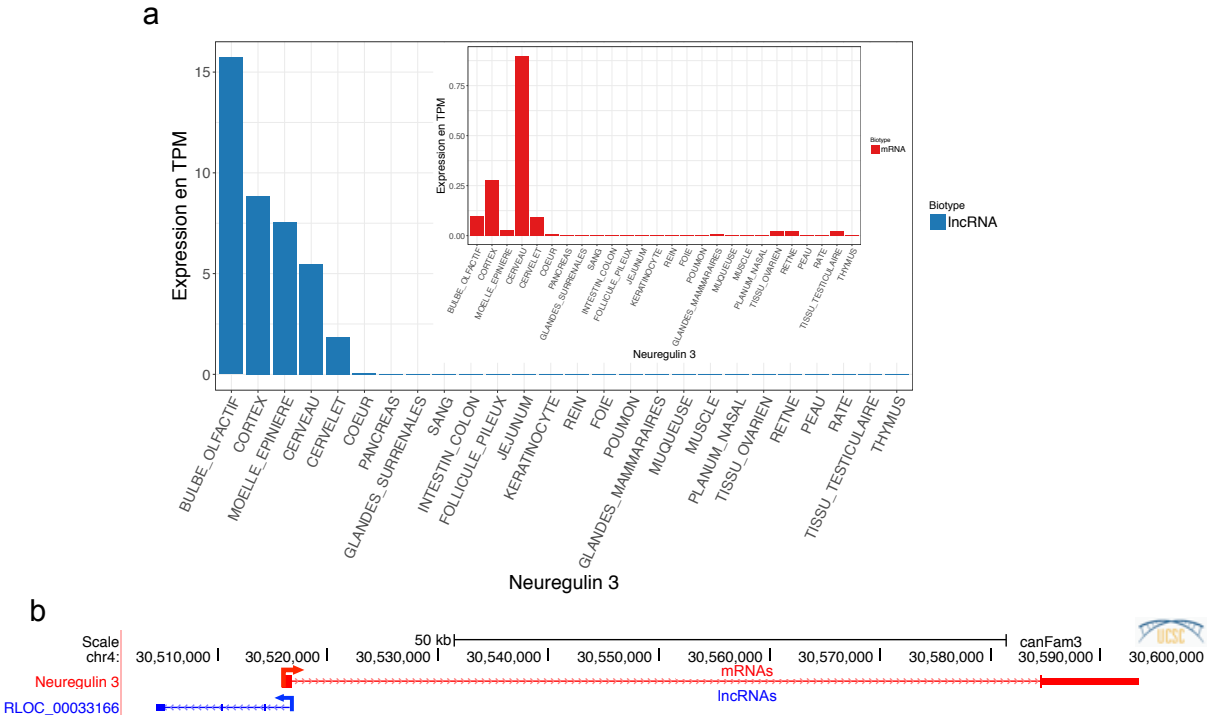

***Figure 16 : Distribution du niveau d'expression de NRG3 et son antisense**, dans les 26 tissus. LncRNA (bleu), antisense du mRNA (rouge) NRG3.*

La classification partitionne les données d'expression en deux principaux clusters qui contiennent plus de cinq tissus. Un premier cluster regroupe les tissus du système nerveux à savoir le bulbe olfactif, le cerveau, le cervelet, le cortex et la moelle épinière et un second cluster regroupe les tissus du système tégumentaire incluant les tissus du follicule pileux, de la glande mammaire, de la kératinocyte, de la muqueuse, de la peau et du planum nasal. Ces clusters reflètent ainsi une certaine proximité histologique des tissus regroupés. L'analyse permet de définir également deux plus petits clusters qui regroupent d'une part les tissus musculo-squelettiques (cœur et muscle) et d'autre part les tissus de l'appareil digestif (intestin_côlon et jéjunum) (Figure 15a.). L'inspection du dendrogramme permet également de déterminer les tissus qui ne clustérisent pas avec les autres tissus. Ainsi, le dendrogramme met en évidence que le tissu testiculaire est isolé des autres, représenté par une branche à part dû à un coefficient de corrélation faible. Les résultats illustrés par la heatmap suggèrent que les lncRNAs retrouvés au sein d'un même cluster, peuvent posséder une fonction commune ou sont exprimés au sein d'un 'pathway' commun au groupe de tissus identifié. La disponibilité des données d'expression pour l'ensemble des gènes, nous permet également de réaliser cette analyse avec les données d'expression des mRNAs. Nous observons les mêmes regroupements de tissus, comprenant un cluster des tissus des systèmes nerveux et un cluster des tissus tégumentaires (Figure 15b.) alors que le tissu testiculaire demeure isolé.

À partir de cette étude, nous avons isolé deux exemples de spécialisation de gènes dans un groupe de tissus. Le lncRNA (*RLOC_00033166*) transcrit en antisens du gène codant pour une protéine, *NRG3* (*Neuregulin3*) impliqué dans la différenciation cellulaire des neuroblastes (Paterson *et al.*, 2017) est exclusivement exprimé dans les tissus du système nerveux ($TPM_{moyen} = 7,89$) (Figure 16). L'expression du gène *NRG3* est également majoritairement présente dans ces mêmes tissus nerveux. Cette co-expression suggère une expression coordonnée entre *NRG3* et le lncRNA (Cf. §2.2. Analyse de co-expression entre lncRNA et gène-codant pour les protéines). Un autre exemple est illustré avec le lncRNA *RLOC_00020746*, majoritairement exprimé dans le système musculo-squelettique (cœur et muscle), qui est transcrit en antisens du gène *POPDC3* ('*Popeye Domain Containing 3'*) impliqué dans le développement des muscles squelettiques et des tissus cardiaques (Andrée *et al.*, 2000). L'identification de modules ou paires de gènes partageant les mêmes patrons

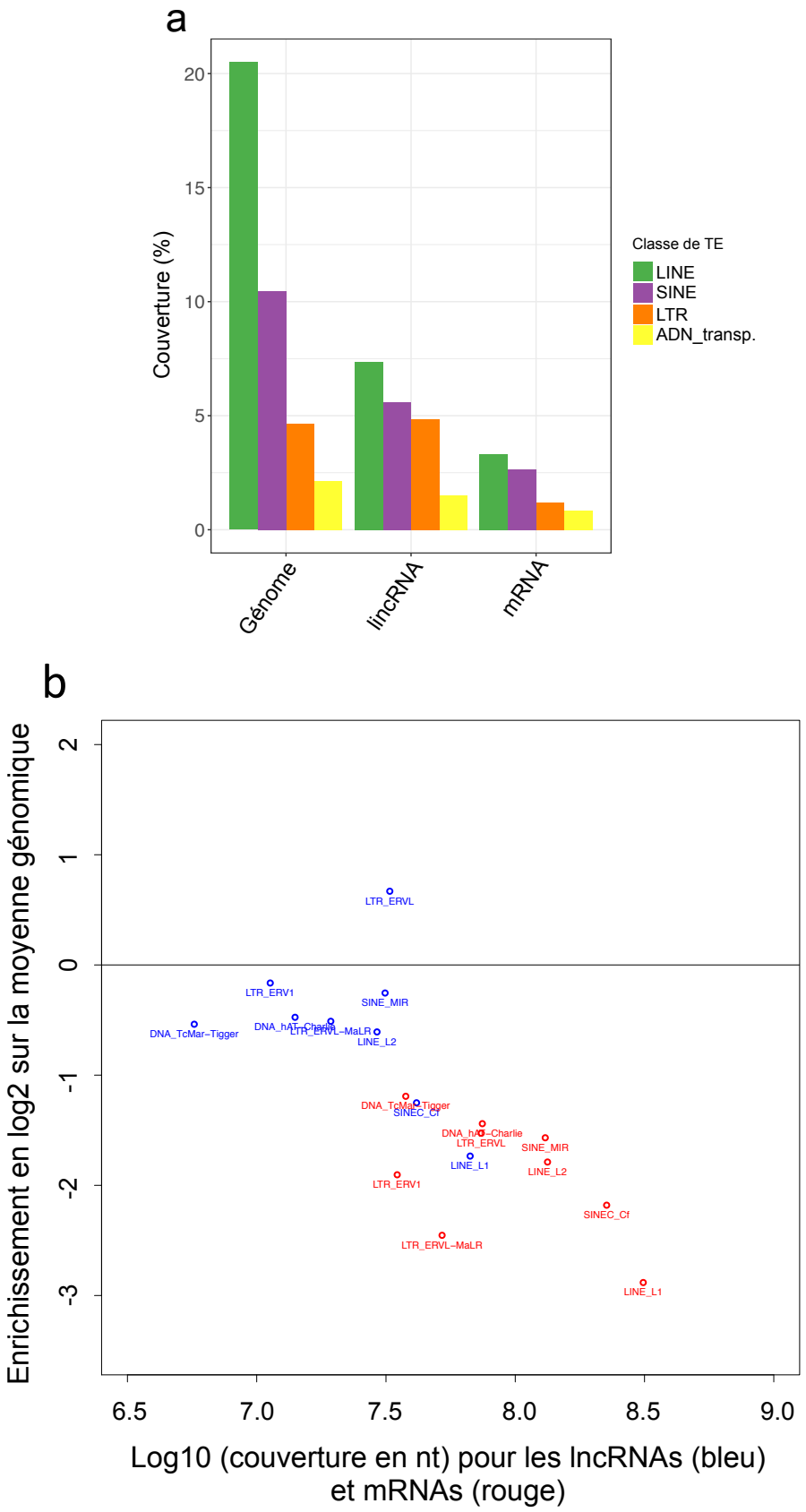

Figure 17 : Répartition et enrichissement des TEs dans le génome. *(a) Proportion d'élément transposable (LINEs, SINEs, LTRs and DNA transposons en vert, violet, orange et jaune respectivement) pour le génome entier du chien, les exons des lincRNAs et les exons des mRNAs (de gauche à droite). (b) Fréquence des familles de TEs pour les lincRNAs (bleu) versus les mRNAs (rouge) comparé à la moyenne génomique. A droite ce trouve les familles les plus grandes. Les enrichissements se trouvent au dessus de l'axe des y et les déplétions en dessous de l'axe des y.*

d'expression suggère de potentielles coordinations ou régulations de l'expression des mRNAs par les lncRNAs.

1.4. Corrélation entre la tissu-spécificité des lncRNAs et leur contenu en Éléments Transposables (TEs)

Des résultats récents ont montré l'importance des TEs dans l'origine (Kelley and Rinn, 2012) et la fonction des lncRNAs (Johnson and Guigó, 2014) et plus particulièrement leur expression tissu-spécifique (Franke *et al.*, 2017). L'identification des TEs canins a été réalisée à l'aide du programme RepeatMasker (Smit, AFA, Hubley, R & Green, 2013), à partir duquel nous avons extrait quatre classes majeures de TEs, les DNA transposons, les 'long interspersed elements' (LINEs), les 'long terminal repeats' (LTRs) et les 'short interspersed elements' (SINEs). Chacune de ces classes est subdivisée en familles, telle que la famille des SINEC_Cf, un SINE spécifique du chien (Cf pour Canis familiaris) (Wang and Kirkness, 2005). Dans cette analyse, nous avons déterminé la proportion de TEs présents au niveau des exons, des régions promotrices définies par les 5 kb en amont du gène et pour la totalité du génome. Nous avons ainsi pu déterminer que près de 70 % des exons des lncRNAs chevauchent au moins un TE. Nous montrons également que près de 20 % des séquences d'exons des lncRNAs sont composées exclusivement de TEs (Figure 17a.). Cette proportion TE-lncRNA est plus faible que celle du génome entier (37,7 %), mais 2,5 fois plus importante que la proportion déterminée pour les exons des mRNAs (7,9 %). Ces résultats mettent en évidence une plus forte présence des TEs au niveau des exons des lncRNAs canins qu'au niveau des exons des mRNAs, comme observé dans d'autres espèces (Kelley and Rinn, 2012). En analysant l'enrichissement en familles de TEs dans les lncRNAs par rapport au génome, nous observons que seule la famille des LTR-ERVL ('Long Terminal Repeat-Endogenous RetroVirus-Like elements') est plus représentée dans les exons des lncRNAs que dans le génome canin (1,6 fois) même si cet enrichissement n'est pas significatif (test de randomisation, p-value > 0,05). Cette famille de TE est également retrouvée enrichie dans les lincRNAs humains d'un facteur de 3,1 (Kelley and Rinn, 2012) et plus généralement, certains auteurs suggèrent que les LTR-ERV sont à la base de la forte transcription observée durant la spermatogénèse (Davis *et al.*, 2017).

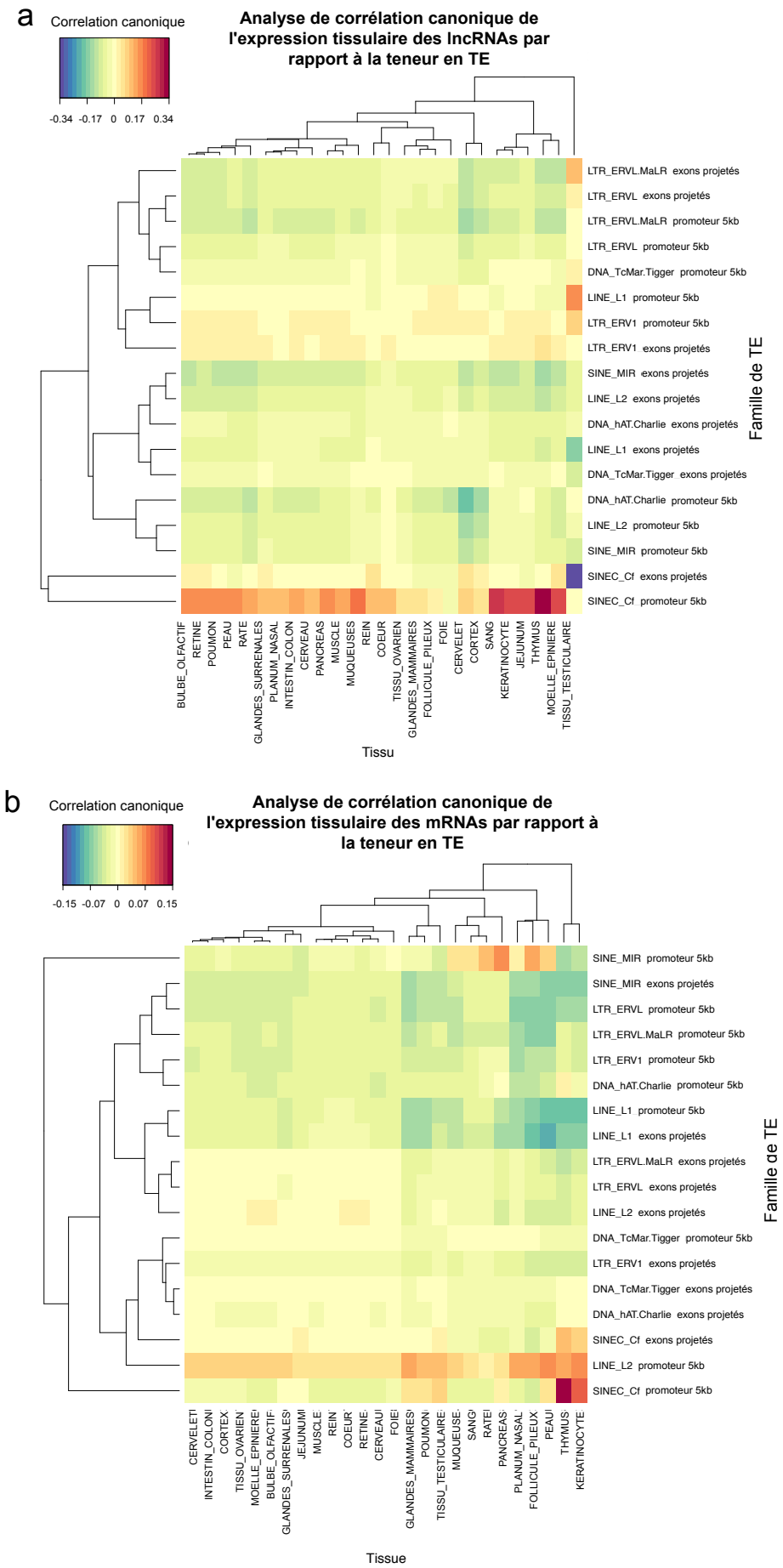

Figure 18 : Corrélations canoniques entre l'expression des gènes dans les 26 tissus *(en colonne) et les familles de TEs dans les exons projetés et 5 kb de séquences en amont des promoteurs (en ligne) ; (a) pour les lncRNAs et (b) pour les mRNAs.*

La présence de TE dans les séquences exoniques et promotrices des lncRNAs suggèrent une relation entre les TEs et l'expression des lncRNAs. Afin de tester cette hypothèse, nous avons réalisé une analyse canonique des corrélations entre le contenu en TE dans les exons (et les régions promotrices) des gènes canins (mRNAs et lncRNAs) et leurs niveaux d'expression en utilisant le programme mixOmics (Rohart *et al.*, 2017) (voir Matériels et Méthodes). Cette analyse met en évidence que l'expression des lncRNAs dans différents tissus tend à être associée à des taux élevés de TEs de type SINEC_Cf dans les régions promotrices (Figure 18a.), une corrélation qui n'est pas observée pour les mRNAs (Figure 18b.). Ces résultats suggèrent que l'insertion de TE, plus particulièrement des éléments mobiles de type SINEC_Cf, dans les séquences promotrices des lncRNAs, peuvent contribuer à une expression localisée spécifique des lncRNAs canins. À l'inverse, cette analyse indique que les exons des lncRNAs exprimés dans le tissu testiculaire sont particulièrement appauvris en SINEC_Cf, suggérant que l'expression spécifique du tissu testiculaire des lncRNAs canins est plus probablement associée à une autre classe de TE comme les LTRs (Figure 18a.) et probablement la famille des LTR-ERVL (Figure 17b.) comme récemment mis en évidence dans la lignée germinale de souris (Franke *et al.*, 2017). Il semble donc que le contenu en LTR-ERV, chez le chien comme chez d'autres mammifères (Davis *et al.*, 2017), soit fortement associé à la forte expression observée dans le tissu testiculaire. À l'inverse pour les autres tissus, la présence de SINEC-Cf dans les promoteurs est associée à la tissue-spécificité. L'analyse de ces corrélations nous permet d'émettre l'hypothèse que, par rapport aux mRNAs, le contenu en TEs des lncRNAs suggèrent leur spécialisation tissulaire en introduisant par exemple, des sites de fixation à des facteurs de transcription spécifiques (TFBS) comme mis en évidence avec l'insulateur CTCF ('CCCTC-binding factor') (Schmidt *et al.*, 2012).

2. Prédiction fonctionnelle bioinformatique des lncRNAs canins

2.1. Analyse des relations d'orthologie des lncRNAs entre l'homme et le chien

2.1.a. La génomique comparative

Entre espèces, deux gènes sont homologues s'ils descendent d'un même gène ancestral, et une analyse comparative de leurs séquences d'ADN permet généralement de

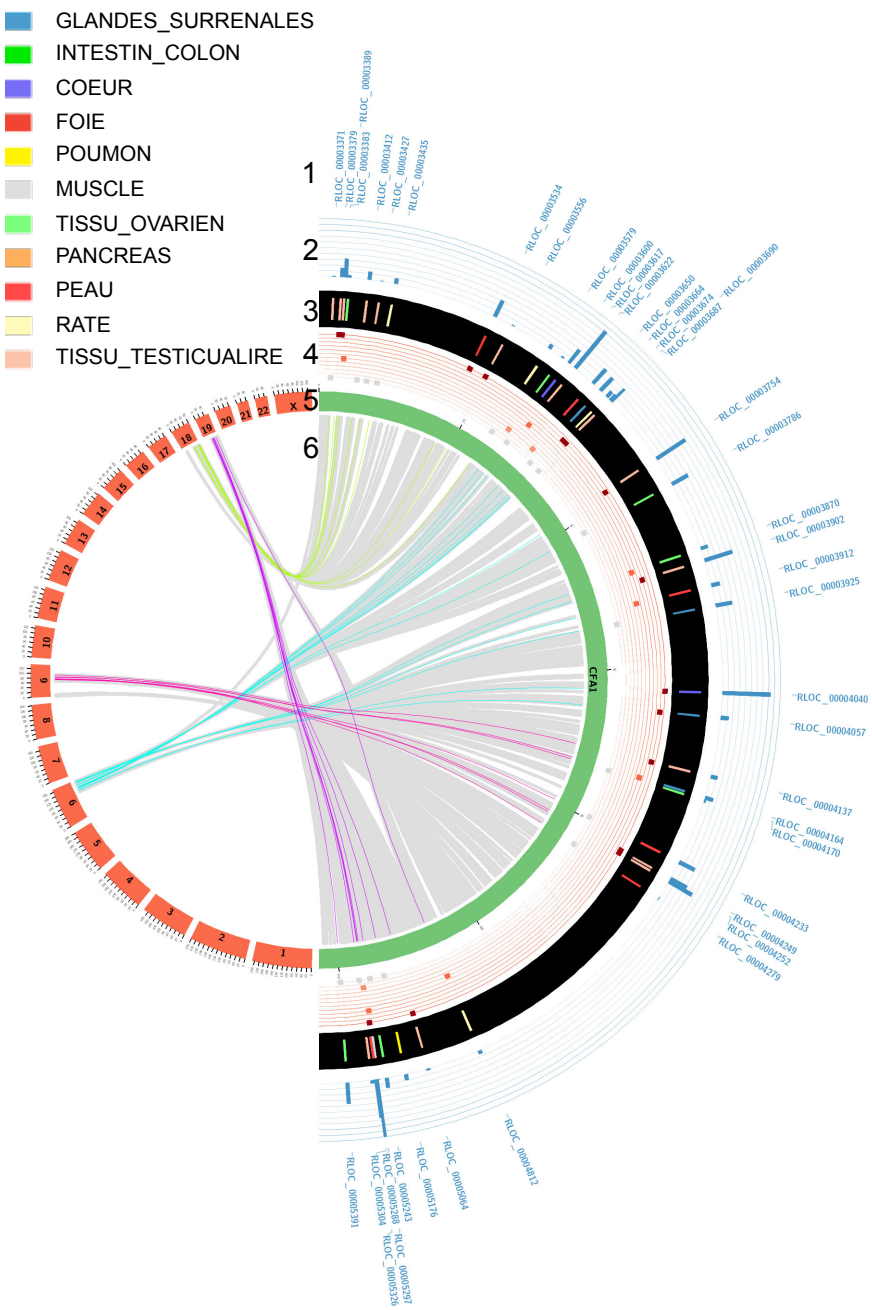

Figure 19 : Diagramme Circos fournissant la visualisation des relations orthologues humaines des lncRNAs et de leurs modèles d'expression transcriptomique comparée. *Pour le chromosome 1 canin, les traces sont décrites de l'extérieur vers l'intérieur. Couche 1 : Des étiquettes des lncRNAs canins identifiés avec des orthologues humains sont décrites sur la piste la plus extérieure de la figure. Couche 2 : Le niveau d'expression des lncRNAs canins est indiqué par l'histogramme bleu. Couche 3 : Le tissu dans lequel le lncRNA est exprimé chez le chien. Les 11 tissus sont la glande surrénale, l'intestin-côlon, le cœur, le foie, le muscle, tissu ovarien, le pancréas, la peau, la rate et le tissu testiculaire et sont représentés avec le code de couleur indiqué dans la légende. Couche 4 : L'expression des lncRNAs orthologues humain est représentée. 11 lignes sont signalées. Lorsque l'orthologue humain lncRNA est exprimé dans le même tissu et avec la plus haute expression, il est représenté par un carré rouge foncé sur la ligne supérieure. Lorsqu'un lncRNA humain orthologue est exprimé dans le même tissu, avec la deuxième expression du maximum d'abondance, il est représenté sur la deuxième ligne en rouge plus clair. Quand un lncRNA humain orthologue n'est pas exprimé dans le même tissu, il est représenté par un carré gris. Couche 5 : La bande verte représente le chromosome canin, la bande rouge représente les chromosomes humains. Couche 6 : Pour la partie intérieure, des lignes colorées relient les relations d'orthologies des lncRNAs entre le chien et l'homme. Les lignes grises représentent les relations d'orthologies des gènes codant les protéines.*

les identifier. Deux gènes sont dits orthologues s'ils ont été séparés par un événement de spéciation et la plupart des gènes orthologues assurent des fonctions conservées entre les espèces. L'identification de la relation d'orthologie permet de transférer les propriétés et les fonctions d'un gène entre espèces. L'identification de gènes orthologues entre espèces est ainsi une approche efficace et pertinente pour produire des annotations fonctionnelles des génomes nouvellement produits et analysés (Herrero *et al.*, 2016).

Cependant la faible conservation de la séquence primaire des lncRNAs rend difficile la découverte d'orthologues lncRNA entre génomes (Ulitsky *et al.*, 2011). Ces problèmes limitent de fait les analyses des origines évolutives, de la dynamique et des fonctions des lncRNAs. La faible conservation de séquence ne suffisant pas à déterminer de manière efficace et exhaustive la conservation des lncRNAs entre espèces, nous avons associé l'analyse de séquence et l'utilisation de la synténie pour rechercher les lncRNAs conservés entre le chien et l'homme. En utilisant la base de données EnsEMBL Compara (Herrero *et al.*, 2016), qui permet d'annoter des régions orthologues entre deux espèces (ou plus), basée sur l'alignement multiples de génomes entiers, nous avons identifié des blocs de synténies conservés entre le chien et l'homme (Figure 19) (voir Matériels et Méthode) puis transposé les coordonnées génomiques des gènes canins sur le génome humain. Nous avons ainsi identifié 939 lncRNAs conservés avec une relation d'orthologie de type 1:1, c'est-à-dire un seul lncRNA humain est identifié orthologue d'un seul lncRNA canin retrouvé dans la région synténique. Les 939 lncRNAs (9 % du total des lncRNAs canins) correspondent aux deux sous-classes suivantes : 560 lincRNAs (9,9 % de la totalité des lincRNAs) et 379 antisens (7,9 % de la totalité des antisens). Les modèles de gène des lncRNAs orthologues identifiés présentent des structures similaires entre le chien et l'homme, avec un nombre moyen d'exon par gène de 3,1 chez le chien et de 3,6 pour l'homme. L'annotation des lncRNAs canins orthologues établie à partir de l'annotation humaine GENCODE (Harrow *et al.*, 2012), nous permet de confirmer d'une part l'identification, chez le chien, des lncRNAs *HOTAIRM* (Chen *et al.*, 2017), *MALAT* (Tripathi *et al.*, 2010), *NEAT_1* (Nakagawa *et al.*, 2011) et *PCA3* (Kretschmer and Tilki, 2017) et d'autre part, d'assigner une relation d'orthologie entre le chien et l'homme pour les gènes suivants : la famille des lncRNAs *CASC* (*CASC15*, *CASC17*, *CASC18* and *CASC20*) et *INHBA-AS1* (Ke *et al.*, 2017) ou encore *MEG9* (Hagan *et al.*, 2009).

Tissus	Nb lncRNA canin > 1 TPM	Nb lncRNA canin > 0,1 TPM	Nb lncRNA humain > 1 TPM	Nb lncRNA humain > 0,1 TPM
Cœur	69	220	42	188
Foie	62	219	62	244
Glandes surrénales	127	309	46	170
Intestin_côlon	128	298	22	95
Muscle	57	188	12	68
Pancréas	20	114	13	77
Peau	132	381	19	132
Poumons	123	366	80	252
Rate	112	343	53	194
Tissu ovarien	165	418	30	134
Tissu testiculaire	374	573	107	378

Table 5 : Nombre de lncRNAs exprimés à > 1 TPM et > 0,1 TPM chez le chien et l'homme.

Des approches complémentaires pour identifier des éléments fonctionnels au sein des génomes s'appuient sur des informations de séquences entre espèces en recherchant les régions qui présentent des signes de contrainte sélective. Une approche de détection des éléments contraints est la méthode GERP ('Genomic Evolutionary Rate Profiling') (Davydov *et al.*, 2010), qui calcule des scores de contrainte pour des positions individuelles d'un alignement multiple, puis définit des segments correspondant à un ensemble de nucléotides contigus hautement contraints. Nous avons évalué les contraintes sélectives agissant sur l'ensemble des séquences correspondant aux lncRNAs et aux mRNAs canins en utilisant les scores GERP disponibles sur UCSC (voir Matériels et Méthodes). Par comparaison, nous montrons que les lncRNAs conservés entre le chien et l'homme tendent à être plus contraints ($GERP_{moyen}$ = 0,221) que les lncRNAs non-conservés ($GERP_{médian}$ = 0,195) et par ailleurs, les 939 lncRNAs conservés sont significativement moins contraints que les mRNAs (test de Wilcoxon, p-value = $9,54^{e-171}$). Bien que statistiquement non significatifs (test de Wilcoxon, p-value = 0,19), ces résultats sont en accord avec d'autres études qui ont rapporté une contrainte sélective intermédiaire agissant sur les lncRNAs (Ulitsky, 2016) par rapport au niveau moyen du génome et par rapport aux valeurs élevées des mRNAs.

2.1.b. Apport de la transcriptomique comparative

L'analyse comparée des génomes s'évalue non seulement par le niveau de conservation de séquences entre espèces mais également par le niveau et la localisation de l'expression des gènes, notamment à l'échelle du transcriptome par des analyses de transcriptomique comparée (Breschi, Gingeras and Guigó, 2017). Les comparaisons inter-espèces des transcriptomes sont maintenant réalisables avec un débit et une résolution très importants obtenus grâce aux données RNA-seq. Dans notre étude, nous avons réalisé une analyse de transcriptomique comparée qui a consisté à comparer les profils d'expression des lncRNAs du chien et de l'homme à partir de onze tissus communs (Table 5). Nous avons utilisé onze données RNA-seq chez l'homme issues du projet ENCODE (ENCODE Project Consortium, 2012) (voir Matériels et Méthodes) incluant la glande surrénale, l'intestin_côlon, le cœur, le foie, le poumon, le muscle, le tissu ovarien, le pancréas, la peau, la rate et le tissu testiculaire.

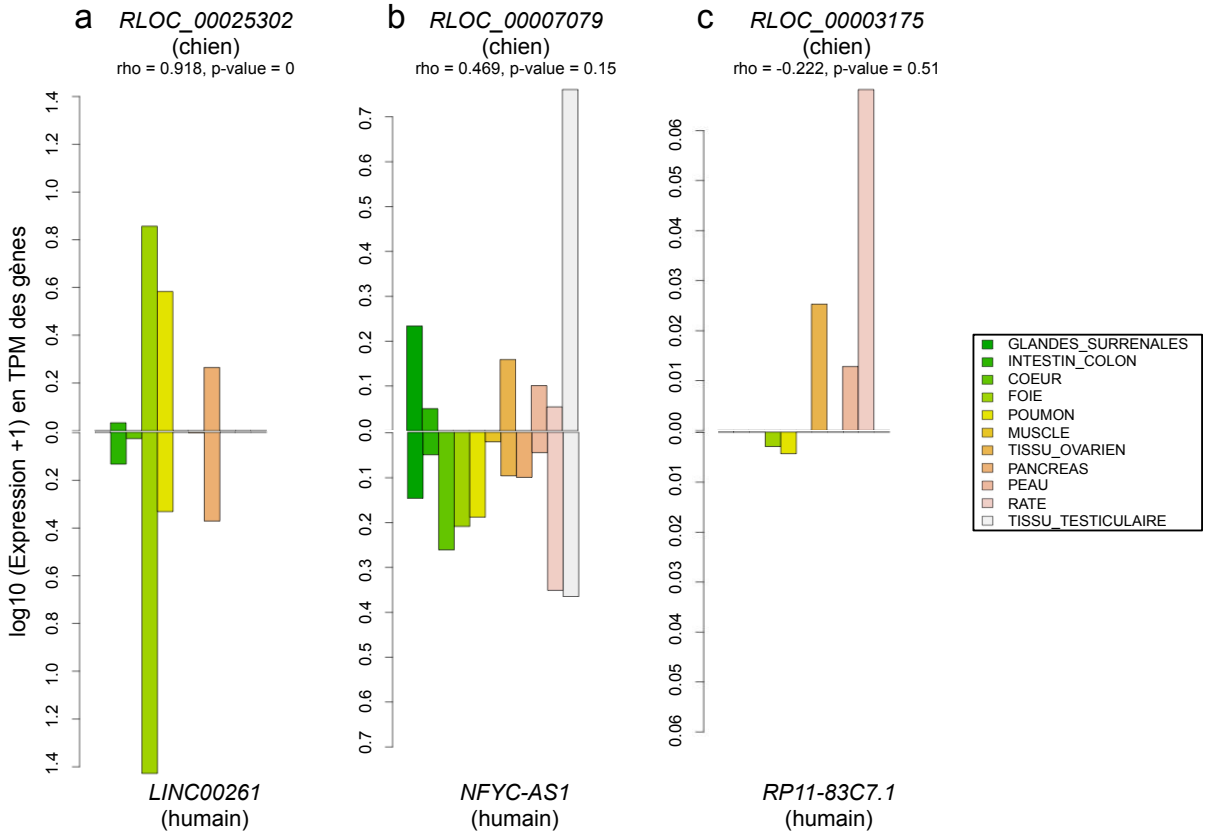

Figure 20 : Diagrammes à barres représentant l'expression des 11 tissus pour le chien et l'homme. *(a) exemple d'un modèle d'expression significativement similaire, (b) exemple d'un modèle d'expression variable, (c) exemple de modèle d'expression différent.*

À partir d'analyses uniformisées des données d'expression entre le chien et l'homme (voir Matériels et Méthodes), nous identifions 727 lncRNAs sur les 939 lncRNAs conservés entre le chien et l'homme, exprimés dans au moins une des deux espèces. L'analyse de corrélation des profils d'expression entre le chien et l'homme indique que les lncRNAs canins tendent à être exprimés dans les mêmes tissus chez l'homme (rho_{moyen} = 0,39) (Figure 20). Cependant, nous avons identifié trois sous-ensembles distincts (Figure 20), avec tout d'abord, la caractérisation de 192 lncRNAs exprimés avec un profil significativement similaire entre le chien et l'homme (rho_{moyen} = 0,87 ; p-value < 0,05), suggérant que ces lncRNAs conservés au niveau de leurs séquences et présentant un même patron d'expression par tissu, possèderaient des fonctions conservées entre espèces (Figure 20a). Ces 192 lncRNAs présentent un fort niveau de tissu-spécificité, tau_{moyen} = 0,94 et 0,93, chez le chien et l'homme respectivement, comme cela avait déjà été rapporté entre l'homme et la souris (Breschi, Gingeras and Guigó, 2017). Ces lncRNAs sont principalement exprimés dans le tissu testiculaire (n = 154), dans la rate (13), le poumon (9), le foie (8), la peau (5) et le cœur (4) (Figure 20a). Dans cette analyse, le lncRNA *TRDN-AS1*, transcrit en antisens du gène *TRDN (triadine*), est spécifiquement exprimé dans le tissu cardiaque à la fois chez l'homme et le chien. Le lncRNA *LINC01698* présente également un profil d'expression et une localisation d'expression commune dans la peau chez le chien et chez l'homme, comme décrit sur le portail de données GTEx (GTEx Consortium, 2013). Par ailleurs, nous identifions un groupe de lncRNAs (n = 77), possédant des profils d'expression corrélés mais avec plus de variabilité d'expression entre les espèces (p-value < 0,2 ; rho_{moyen} = 0,50) (Figure 20b). Ce sous-ensemble montre que l'identification d'une relation d'orthologie n'implique pas une expression exactement conservée et que différents facteurs chez les deux espèces peuvent agir sur l'expression du même gène et moduler son niveau d'expression. Un troisième sous-ensemble de lncRNAs (n = 458) a été défini avec des profils transcriptionnels relativement divergent entre le chien et l'homme, qui ainsi séparent les espèces plus que les systèmes anatomiques (valeur p-value > 0,2 ; rho_{moyen} = 0,17) (Figure 20c). Ce sous-ensemble peut réunir des lncRNAs exprimés dans des systèmes anatomiques avec des structures complexes composées de nombreux types de cellules primaires, pour lesquels la fonctionnalité tissulaire peut différer entre les échantillons et/ou entre les espèces en raison de leur structure. Ainsi le lncRNA canin (*RLOC_00028824*), orthologue du lncRNA humain

CTC_232P5.1, présente un profil d'expression totalement différent de celui l'homme. Sur les 11 tissus représentés dans cette étude, le lncRNA s'exprime entre 0,47 TPM dans la peau et 164,94 TPM dans le tissu testiculaire (TPM_{moyen} = 22,34 TPM) et n'est pas exprimé dans seulement deux tissus. Contrairement à son orthologue humain présentant une expression extrêmement faible et dans un unique tissu, à savoir le foie (0,052 TPM).

2.2. Analyse de co-expression entre lncRNAs et mRNAs canins

Dans plusieurs études, il a été démontré que les lncRNAs sont impliqués dans la régulation de l'expression en coordonnant différents mécanismes de la transcription, des processus épigénétiques, ou post transcriptionnels (Mallory and Shkumatava, 2015). L'analyse de la co-expression de paires de gènes dans un même environnement génomique est un moyen d'identifier des séquences qui potentiellement modulent la transcription des gènes voisins. De plus, les séquences régulatrices sont souvent situées à proximité de leur gène cible, modulant leur transcription. Par conséquent, l'identification de paires co-exprimées de lncRNAs et de mRNAs fournit un moyen (i) d'identifier des relations entre les lncRNAs et les mRNAs et (ii) d'inférer la fonctionnalité régulatrice potentielle des lncRNAs. L'analyse de co-expression a été réalisée avec la stratégie 'guilt-by-association' (Guttman *et al.*, 2009), qui émet l'hypothèse que lorsqu'un lncRNA est proche au niveau génomique et co-exprimé avec un mRNA, une association fonctionnelle entre les deux éléments est plus probable. Le module 'classifier' de l'outil FEELnc (Wucher *et al.*, 2017) a été utilisé afin d'obtenir l'ensemble des paires de lncRNA:mRNA dans une fenêtre de 1 Mb autour de chacun des lncRNAs (voir Matériels et Méthodes). Nous avons identifié plus de 126 000 paires, que nous avons assigné à deux classes, d'une part les paires antisens:mRNA (n = 3 401) et d'autre part les paires lincRNA:mRNA (n = 123 456), définies par les lncRNAs intergéniques localisés à moins d'1 Mb du mRNA.

En utilisant les données d'expression provenant des 26 RNA-seq, nous avons calculé la corrélation de Spearman entre le lncRNA et le mRNA de chaque paire. Au total, après avoir filtré selon les critères suivants (| rho | > 0,5 ; p.ajustée BH (Cai and Chan, 2016) < 0,05), 7 615 paires lincRNA:mRNA et 524 paires antisens:mRNA ont été identifiées comme significativement corrélées, correspondant à un sous-ensemble de 3 410 lncRNAs uniques. Parmi ces paires, 58 étaient négativement corrélées, pour lesquelles nous avons déterminé

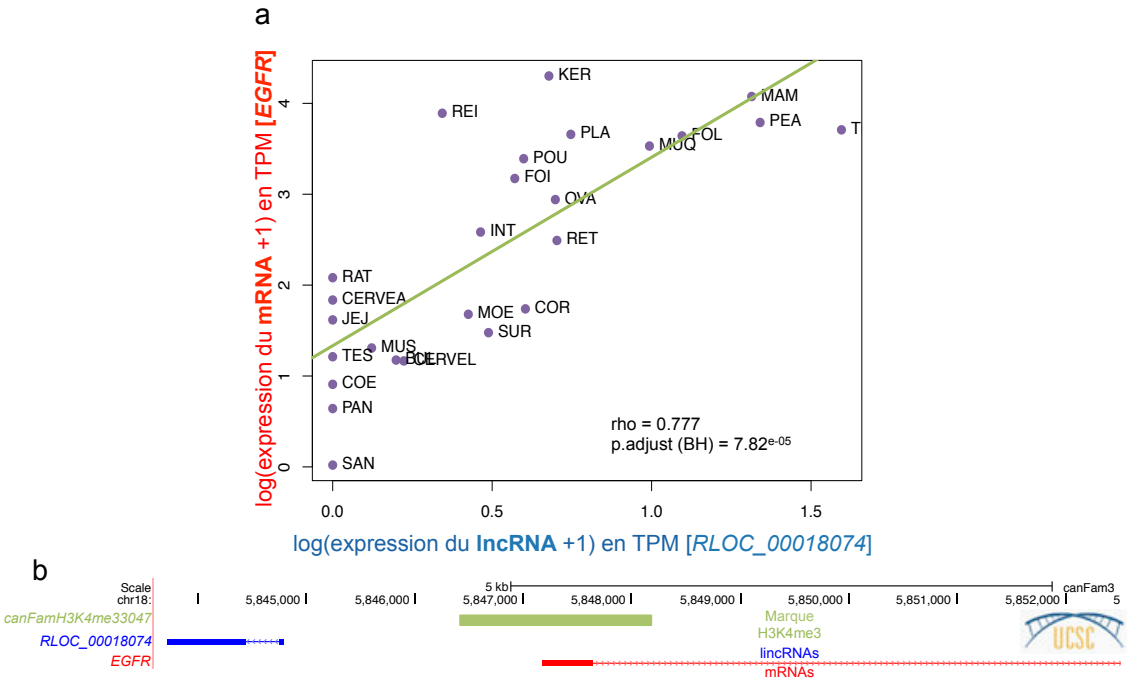

Figure 21 : Corrélation de la paire lncRNA:EGFR. *(a) Diagramme représentant la corrélation d'expression de la paire lncRNA:EGFR, les 26 tissus sont représentés. (b) Capture d'écran UCSC de la paire lncRNA:EGFR, avec une marque d'histone H3K4me3.*

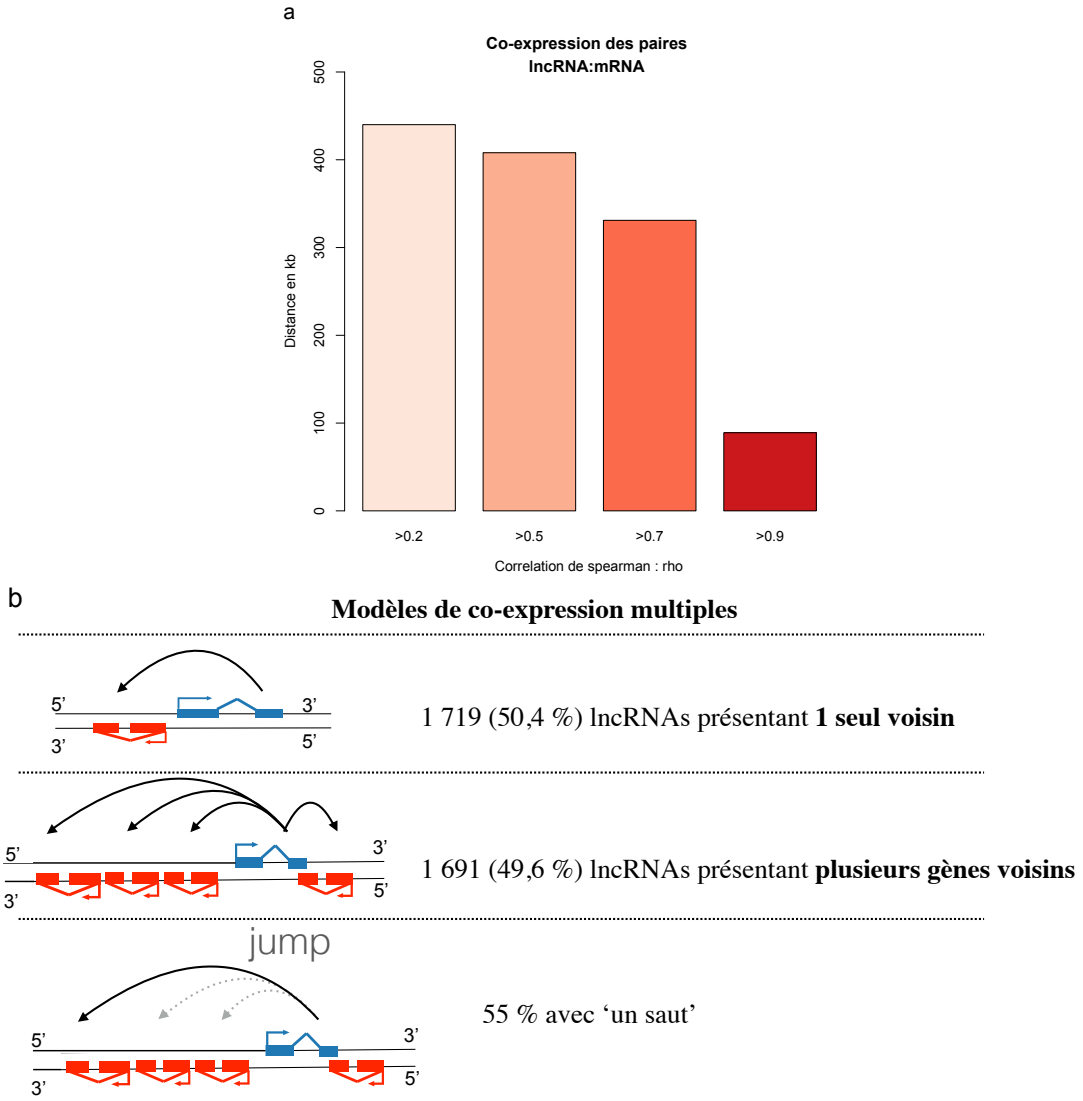

Figure 22 : Schémas de co-expression des paires la paire lncRNA:mRNA. *(a) lncRNA:mRNA co-expression, valeurs de Spearman rho de > 0,2 à > 0,9 en fonction de la distance entre le lncRNA et le mRNA. (b) Schéma représentant plusieurs modèles de co-expression, avec seulement 1 voisin, plusieurs gènes voisins et un saut.*

une distance moyenne entre les lncRNAs et leurs cibles mRNAs, de 471 kb, contre 394 kb pour toutes les paires positivement corrélées (test de Wilcoxon, unilatéral p-value = 0,02). Parmi cet ensemble de paires lncRNA:mRNA avec une expression corrélée significative, nous avons extrait une corrélation positive entre le lincRNA (*RLOC_00018074*) et le mRNA *EGFR* (‘*Epidermal Growth Factor Reception’*) impliqué dans des mécanismes d’oncogenèse et de migration tumorale (Kerboua *et al.*, 2016) (Figure 21a). De plus, la présence d’une marque épigénétique de triméthylation de la lysine 4 de l'histone 3 (H3K4me3), entre les sites d'initiation de la transcription des lincRNA et l’*EGFR* suggère la présence d’un promoteur potentiellement bi-directionnel, (Villar *et al.*, 2015) (Figure 21b), ce qui renforce l’hypothèse d’une relation fonctionnelle entre le lincRNA *RLOC_00018074* et *EGFR.*

2.2.a. Distance et type de la relation de co-expression

Nous avons observé plus de lncRNAs antisens (36,5 %) impliqués dans des paires lncRNA:mRNA significativement co-exprimées que de lincRNAs (29,4 %) (test du Khi^2, p-value < $5{,}5^{e-8}$). Cette différence est probablement due au fait que les antisens chevauchent les mRNAs, favorisant une expression conjointe de ces gènes dans chacun des tissus. Parmi les 3 410 lncRNAs co-exprimés avec un mRNA, nous mettons en évidence plusieurs types de relation de co-expression. Ainsi 50,4 % des lncRNAs sont co-exprimés avec un seul mRNA et donc 49,6 % avec plusieurs mRNAs (de 2 à 25) dans l’intervalle de 1 Mb. Nous avons caractérisé la relation entre la distance et la corrélation des paires co-exprimées. Les paires très corrélées (0,9 < | rho | ≤ 1) sont plus proches ($distance_{moyenne}$ = 89 kb) que les paires de gènes co-exprimés avec une valeur rho inférieure (0,5 < | rho | < 0,7 ; $distance_{moyenne}$ = 411 kb) (Figure 22a). Comme décrit dans d’autres études (Tsai *et al.*, 2009), ceci indique que les paires de gènes co-exprimées conservent un fort niveau de proximité au cours de l'évolution et que la plupart des co-expression en *cis* peuvent être identifiées au sein d’un intervalle génomique de 1 Mb. De manière intéressante, nous identifions pour plus de 800 paires (55 % des lncRNAs présentant plusieurs voisins mRNAs), que le lncRNA présente une corrélation plus élevée avec le deuxième (ou plus) mRNA le plus proche (Figure 22b). Ces résultats confirment que la co-expression n’est pas entièrement corrélée avec la distance génomique, et suggèrent que d’autres facteurs tels que l’organisation spatiale de la chromatine dans la cellule pourraient corréler avec la co-expression. En effet,

	Description	FDR	Nb de gène
GO:0007423	'Sensory organ development'	$0,6^{e-3}$	147
GO:0007389	'Pattern specification process'	$0,7^{e-3}$	127
GO:0048736	'Appendage development'	$4,1^{e-3}$	58
GO:0048089	'Positive regulation of cellular component biogenesis'	$4,1^{e-3}$	129
GO:0061564	'Axon development'	$5,1^{e-3}$	129
GO:0018205	'Peptidyl lysine modification'	$5,1^{e-3}$	115
GO:0051961	'Negative regulation of nervous system development'	$6,7^{e-3}$	81
GO:0030902	'Hinbrain development'	$6,7^{e-3}$	49
GO:0010975	'Regulation of neuron projection development'	$7,5^{e-3}$	116
GO:0048568	'Embryonic organ development'	$8,9^{e-3}$	118
GO:0050804	'Modulation of synaptic transmission'	$8,9^{e-3}$	84
GO:0016049	'Cell growth'	$1,1^{e-2}$	135
GO:0007033	'Vacuole organization'	$1,1^{e-2}$	54
GO:0016570	'Histone modification'	$1,1^{e-2}$	120
GO:0060021	'Palate development'	$1,2^{e-2}$	32
GO:0060560	'Developmental growth involved in morphogenesis'	$1,7^{e-2}$	65
GO:0051962	'Positive regulation of nervous system development'	$1,9^{e-2}$	121
GO:0001701	'In utero embryonic development'	$2,0^{e-2}$	92
GO:0022604	'Regulation of cell morphogenesis'	$2,2^{e-2}$	118
GO:0016358	'Dendrite development'	$2,4^{e-2}$	58
GO:0007224	'Smoothened signaling pathway'	$3,5^{e-2}$	40
GO:0010720	'Positive regulation of cell development'	$4,0^{e-2}$	118

Table 6 : Analyse d'enrichissement de termes GO *mRNA significativement co-exprimés dans 8 139 paires lncRNA:mRNA en utilisant le site web WebGestalt (www.webgestalt.org/). Les termes GO sont classés en fonction de leur FDR avec le nombre de gènes présentant le terme GO spécifique.*

la connaissance de l'organisation spatiale des chromosomes permet de mieux comprendre les interactions entre loci distants dans un espace en trois dimensions.

2.2.b. Caractérisation par ontologies de gènes des lncRNAs impliqués dans les co-expressions.

L'utilisation des ontologies de gènes est régulièrement utilisée pour préciser les relations entre groupes de gènes, à partir de données descriptives qui sont associées pour produire de la connaissance dans l'annotation fonctionnelle des génomes. À partir des résultats de co-expression lncRNA:mRNA, nous avons réalisé une étude d'enrichissement des termes GO ('Gene Ontology') associés aux mRNAs conservés entre le chien et l'homme (voir Matériels et Méthode). Ainsi l'utilisation de annotation fonctionnelle des mRNAs permet de prédire une fonctionnalité potentielle des lncRNAs co-exprimés. L'analyse GO (voir Matériels et Méthodes) a permis de mettre en évidence 22 termes GO significativement enrichis (p.ajustée BH < 0,05, Table 6) correspondant à des processus du développement tels que 'sensory organ development' (GO: 0007423, 147 gènes), 'axon development (GO: 0061564, 129 gènes) ou 'hindbrain development' (GO: 0030902, 49 gènes). Ces 22 termes GO ont été mis en évidence à l'aide de 2 920 mRNAs conservés et co-exprimés avec plus de 900 lncRNAs canins. Ces résultats prédisent une attribution fonctionnelle à ces lncRNAs canins et représentent une approche '*in silico*' pour inférer un rôle potentiel aux lncRNAs.

III. Implication des lncRNAs dans le mélanome oral canin

Dans le cadre d'un projet de recherche sur le mélanome, nous proposons le chien comme modèle spontané du mélanome oral, une forme agressive de mélanome. Le mélanome oral est relativement fréquent chez le chien avec près de 62 % des mélanomes et l'équipe Génétique du chien a collecté et caractérisé depuis plusieurs années une cohorte d'échantillons tumoraux et contrôles (Gillard *et al.*, 2014). À partir d'un ensemble de 60 échantillons dont 46 échantillons tumoraux et 14 échantillons appariés pour lesquels nous disposons du prélèvement d'une fraction tumorale et du prélèvement de tissu contrôle pour un même individu, nous avons investigué les profils d'expression de l'ensemble des lncRNAs en se basant sur des données de séquençage du transcriptome (RNA-seq). L'analyse et l'identification des changements d'expression génique dans le cancer apportent des

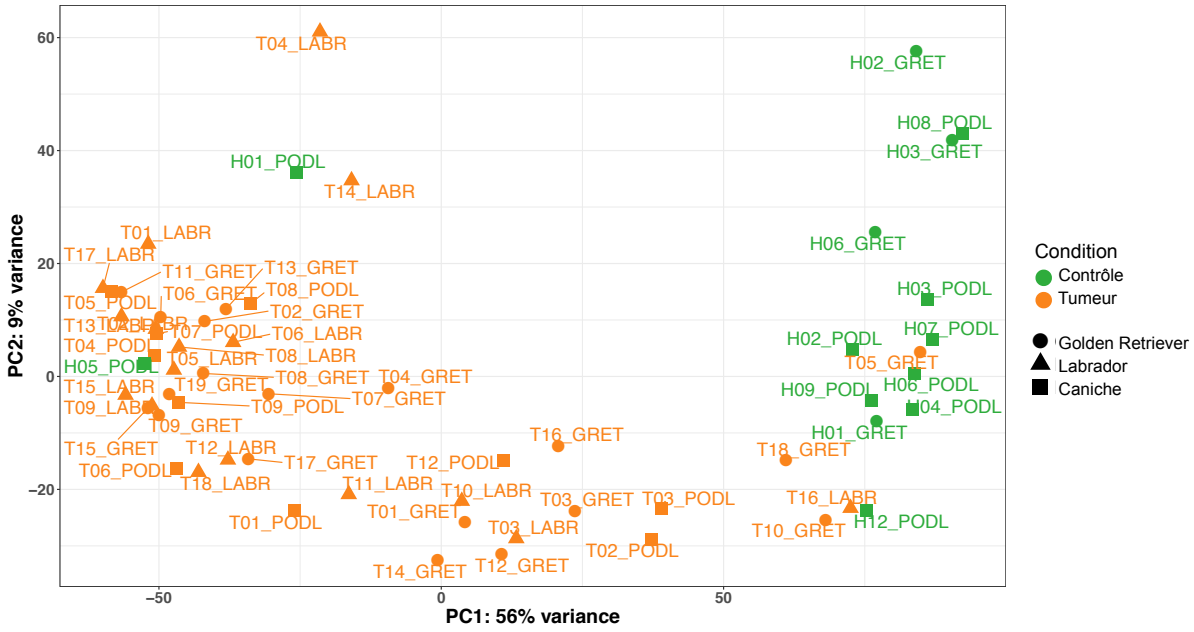

Figure 23 : Analyse en composante principale des 60 échantillons initiaux de tumeurs de mélanome oral canin *(en orange) et leurs contrôles en conditions appariés pour 14 d'entre elles (en vert).*

opportunités pour identifier les biomarqueurs qui peuvent être informatifs pour l'établissement de diagnostics et le choix des options de traitement.

L'analyse des profils de transcription a été réalisée par le pipeline défini au laboratoire, incluant l'étape de mapping (STAR), et l'étape de l'évaluation de l'abondance de chaque gène par le programme RSEM (voir Matériels et Méthodes). Après normalisation des données de comptages, nous avons réalisé une analyse de différentiel d'expression pour caractériser les variations significatives d'expression entre deux conditions, ici, la condition tumorale et la condition contrôle. Cette analyse nous permet ainsi d'identifier des lncRNAs candidats pour étudier la régulation de la progression tumorale des mélanomes chez le chien mais pourront également être évalués pour leurs potentiels diagnostics et en thérapeutiques en médecine humaine.

1. Contrôle qualité des échantillons de mélanomes oraux

Une analyse en composante principale (ACP) a tout d'abord été réalisée, afin de caractériser le regroupement de l'ensemble des échantillons via les profils d'expression des 10 444 lncRNAs (Figure 23). La représentation graphique issue de cette analyse, permet d'observer que l'axe 1 explique 56 % de la variance et semble discriminer les échantillons tumoraux (à gauche) des échantillons contrôles (à droite). L'interprétation de l'axe 2 (9 % de la variance) peut-être liée à plusieurs effets, tels que l'hétérogénéité des échantillons, la présence de plusieurs types cellulaires ou encore un possible effet race. Pour l'axe 1, cette analyse descriptive de la qualité des données met en évidence que deux échantillons contrôles, H01_PODL (échantillon contrôle du Caniche n°1) et H05_PODL (échantillon contrôle du Caniche n°5) se regroupent avec les échantillons tumoraux. Nous avançons plusieurs hypothèses pour expliquer ce regroupement, notamment de possibles artefacts techniques, un prélèvement imprécis de la zone contrôle, une erreur d'étiquetage du prélèvement ou la contamination de la zone contrôle par des cellules tumorales. Pour la suite de cette analyse, nous avons décidé d'exclure ces 2 échantillons contrôles ainsi que leurs échantillons tumoraux correspondants. À l'inverse, quatre tumeurs semblent être regroupées près des échantillons contrôles sur l'axe 1 (T05_GRET, T18_GRET, T10_GRET et T16_LABR), et pour compléter, deux tumeurs apparaissent isolées des autres sur l'axe 2 (T04_LABR et T14_LABR), ces données ne seront donc pas prises en compte. L'analyse de

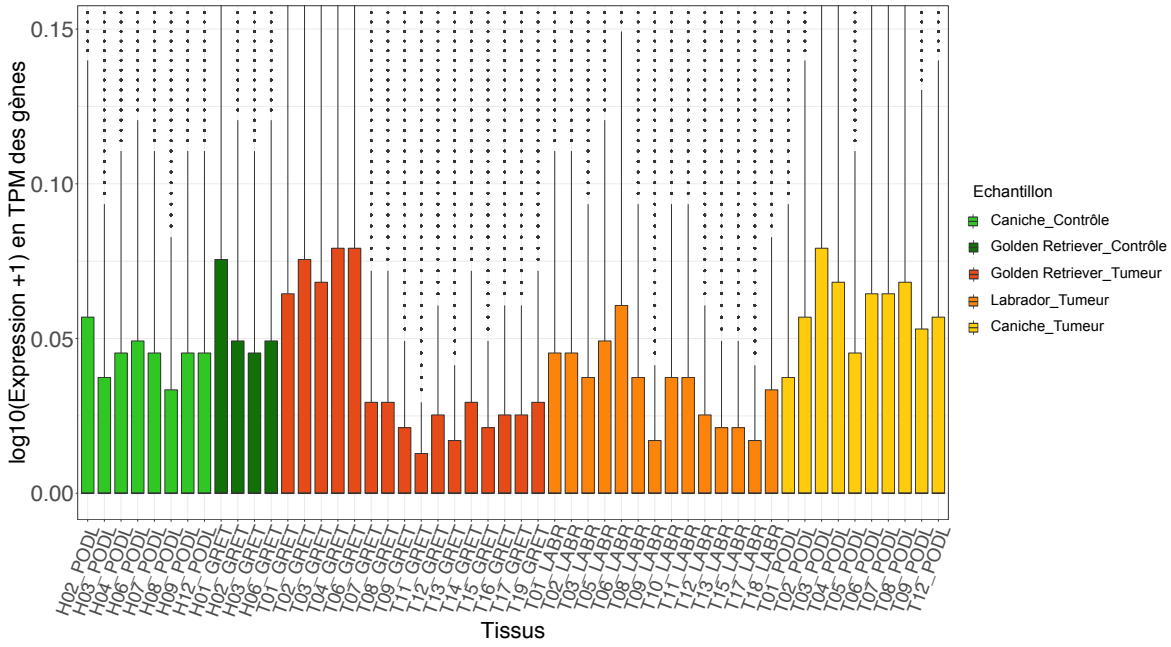

Figure 24 : Analyse des niveaux de transcription des lncRNAs dans les 52 échantillons de tumeurs de mélanome oral canin *et leurs contrôles en conditions appariés pour 12 d'entres eux, en TPM.*

	|FC| > 0	|FC| > 2
Nombre de lncRNAs Différentiellement Exprimés	**a**	**b**
Nombre de lncRNAs Différentiellement Exprimés Et conservés avec l'homme	**c**	**d**

Figure 25 : Diagramme de Venn permettant de visualiser le nombre de lncRNAs différentiellement exprimés (DE) en commun entre ceux issus de DESeq2 (en bleu) et ceux issus d'EBSeq (en rouge). *(a) Nombre de lncRNAs DE avec un | log Fold Change (FC) | > 0. (b) Nombre de lncRNAs DE avec un | lFC | > 2. (c) Nombre de lncRNAs DE avec un | lFC | > 0 et conservé avec l'homme. (d) Nombre de lncRNAs en DE avec un | lFC | > 2 et conservé avec l'homme.*

différentiel d'expression est ensuite réalisée sur un ensemble de 52 échantillons RNA-seq composés de 40 tumoraux et 12 contrôles.

2. Les lncRNAs différentiellement exprimés entre tumeurs et contrôles

L'analyse des profils d'expression est réalisée sur la matrice d'expression en TPM (Transcript Per Million). Cette analyse révèle que 32 % (3 356 / 10 444) des lncRNAs sont exprimés dans au moins un échantillon avec une valeur de TPM > 1, indiquant une bonne représentativité d'expression de notre jeu de données. La Figure 24 nous permet d'observer le profil d'expression de ces 52 échantillons, sur l'ensemble des 10 444 lncRNAs pour lesquels les moyennes de TPM sont comprises entre 0,55 (H08_PODL) et 1,76 (T07_PODL). Néanmoins, nous n'avons pas représenté les valeurs extrêmes ('outliers'), allant jusqu'à 9 233,59 TPM pour l'échantillon T07_PODL.

2.1. DESeq2 versus EBSeq

Afin de réaliser l'analyse de différentiel d'expression, nous avons utilisé deux outils complémentaires, d'une part le programme DESeq2 (Love, Huber and Anders, 2014) et d'autre part le programme EBSeq (Leng *et al.*, 2013). Même si les deux approches font l'hypothèse que les données de comptage suivent une distribution négative binomiale, DESeq2 va tester le différentiel d'expression par un test de Wald alors que EBSeq estime les probabilités a posteriori de l'expression différentielle à l'aide de méthodes empiriques bayesiennes (Gamalo-Siebers *et al.*, 2016). L'analyse est réalisée à partir de données de comptages normalisées de 52 RNA-seq issus des échantillons de 40 tumeurs de 10 Caniches, 16 Golden Retrievers et 14 Labradors, ainsi que de 12 contrôles correspondant à 8 des 10 Caniches et 4 des 16 Golden Retrievers. Pour l'ensemble de l'étude des gènes DE, nous fixons le FDR < 0,01 afin de comparer les outils.

Dans un premier temps, nous identifions les lncRNAs différentiellement exprimés (DE) sans prendre en compte l'intensité du différentiel (*i.e.* log Fold Change ou lFC), et obtenons un total de 1 542 lncRNAs DE pour DESeq2 et 840 lncRNAs pour EBSeq, avec une intersection de plus de 750 lncRNAs DE, soit 31,6 % des lncRNAs caractérisés DE (Figure 25a.). La différence est du simple au double entre les résultats de ces deux outils. Le grand nombre de lncRNAs DE seulement retrouvés par l'outil DESeq2 (n = 790) est assez surprenant sachant

qu'une étude comparative de 2015 a montré un nombre moins important de faux-positifs pour DESeq2 comparativement à EBSeq (Seyednasrollah, Laiho and Elo, 2015). Cette observation peut-être atténuée du fait que l'analyse est réalisée sur un jeu de données de 52 échantillons contre ~ 10 à 15 dans l'étude sur les faux-positifs.

Dans un second temps, nous avons fixé le seuil de log Fold Change à une valeur plus importante avec l'hypothèse biologique sous-jacente que plus le seuil est fort, plus le lncRNA est susceptible de présenter un effet fort dans le progression tumorale. Ce seuil, fixé à | lFC | > 2, a permis d'affiner la liste de lncRNAs à 482 lncRNAs DE en commun entre les deux outils (Figure 25b.), soit 64 % des lncRNAs identifiés DE avec un lFC > 0.

Nous nous sommes ensuite intéressés aux lncRNAs DE et conservés avec l'homme. L'identification de ces lncRNAs impliqués dans des mélanomes muqueux canins et conservés chez l'homme, pourrait permettre la découverte puis l'étude de nouveaux biomarqueurs de gènes impliqués dans le diagnostic des mélanomes humains et canins ou encore être utilisés en thérapie. La caractérisation des orthologues chien-homme réalisée dans la première partie du travail de thèse (voir II.2.2. Analyse de co-expression entre lncRNAs et mRNAs canins), nous permet d'identifier, tout d'abord sans filtre lFC, 107 lncRNAs DE en commun entre l'analyse DESeq2 et EBSeq et aussi conservés avec l'homme (Figure 25c.), correspondant à 11,5 % des 939 lncRNAs identifiés conservés entre le chien et l'homme. Cette liste, affinée grâce à l'application du seuil (| lFC | > 2), nous permet d'obtenir un sous-ensemble de 72 lncRNAs soit 7,7 % de lncRNAs candidats par rapport aux 939 lncRNAs précédemment identifiés comme conservés entre le chien et l'homme (Figure 25d.). Ces 72 lncRNAs, sur ou sous-exprimés dans les tumeurs et conservés entre le chien et l'homme, sont donc de bons candidats en tant que biomarqueur dans le cas du mélanome buccal chez le chien et potentiellement chez l'homme.

2.2. DESeq2 versus $DESeq2_{multifacteur}$

L'outil DESeq2, en plus de tester l'effet condition 'Tumeur' vs 'Contrôle', peut prendre en compte un design plus sophistiqué incluant des effets supplémentaires ou co-variables. Notre jeu de données est composé de trois races différentes et donc il apparaît intéressant de contrôler cet effet race dans l'analyse du différentiel d'expression. Par ailleurs, les données issues d'échantillons muqueux possèdent généralement deux types cellulaires

	| IFC | > 0	| IFC | > 2
Nombre de lncRNAs Différentiellement Exprimés	a (multifacteur)	b (multifacteur)
Nombre de lncRNAs Différentiellement Exprimés Et conservés avec l'homme	c (multifacteur)	d (multifacteur)

Figure 26 : Diagramme de Venn permettant de visualiser le nombre de lncRNAs différentiellement exprimés (DE) en commun entre ceux issus de DESeq2$_{_défaut}$ (en bleu) et ceux issus de DESeq2$_{_multifacteur}$ (en rouge). *(a) Nombre de lncRNAs DE avec un | log Fold Change (FC) | > 0. (b) Nombre de lncRNAs DE avec un | lFC | > 2. (c) Nombre de lncRNAs DE avec un | lFC | > 0 et conservé avec l'homme. (d) Nombre de lncRNAs en DE avec un | lFC | > 2 et conservé avec l'homme.*

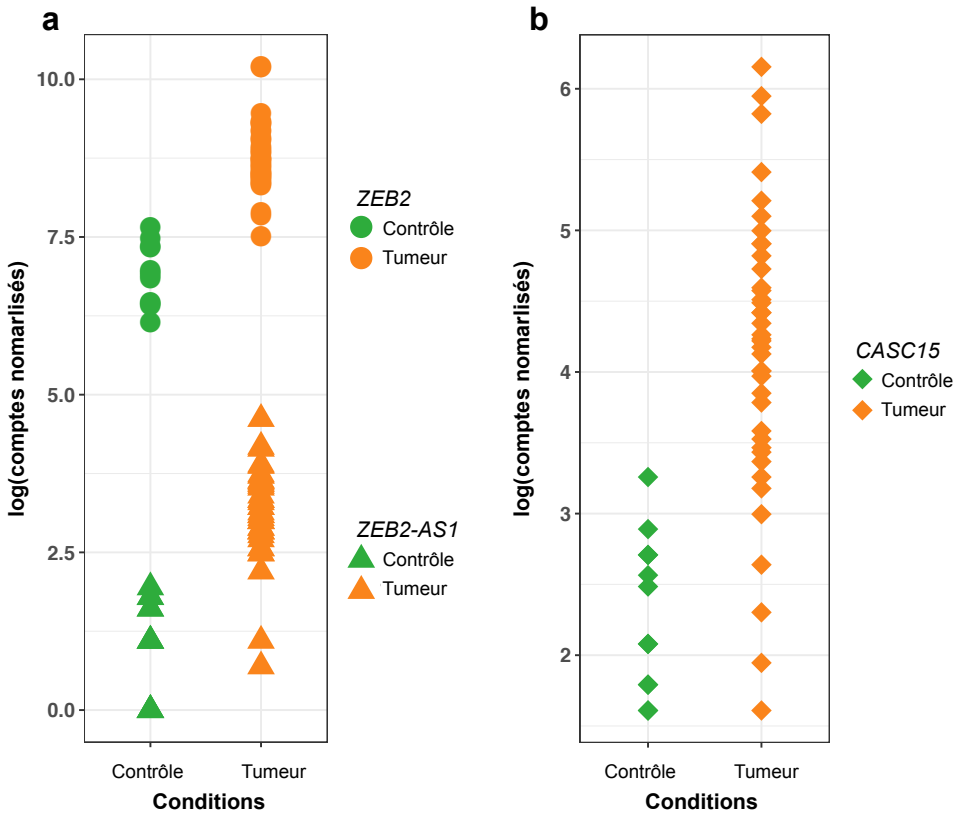

Figure 27 : Deux exemples de gènes DE*, tumeurs de mélanome oral canin (en orange) par rapport aux contrôles (en vert) ; (a) exemple du lncRNA ZEB2-AS1 activant l'expression du mRNA ZEB2, sur-exprimé dans les tumeurs, (b) exemple du lncRNA CASC15 sur-exprimé dans les tumeurs.*

différents, les kératinocytes et ceux qui nous intéresse dans cette étude, les mélanocytes (voir Matériels et Méthodes). Cet effet type cellulaire est également inclus dans ce nouveau design de DESeq2. Pour analyser l'importance de la prise en compte de l'effet race et type cellulaire dans l'identification de gènes DE, nous comparons les résultats de DESeq2 avec les paramètres par défaut $DESeq2_{défaut}$ (c'est-à-dire le test seul de l'effet condition : 'Tumeur' versus 'Contrôle'), versus DESeq2 prenant en compte l'effet race et l'effet type cellulaire ($DESeq2_{multifacteur}$). L'analyse $DESeq2_{multifacteur}$ identifie moins de lncRNAs DE (n = 200) sans filtre lFC (Figure 26a.) suggérant que la prise en compte de l'effet race et type cellulaire permet de réduire le nombre de faux-positifs. Cette analyse permet également de mettre en évidence 66 lncRNAs non identifiés dans l'analyse $DESeq2_{défaut}$. En appliquant le seuil | lFC | > 2, seul 8 % des 200 gènes lncRNAs ne passent pas ce seuil contre 30 % avec les paramètres par défauts (Figure 26b.) ce qui semble signifier que l'analyse $DESeq2_{multifacteur}$ est bien plus conservative pour l'identification des lncRNAs DE.

Comme précédemment, nous avons cherché les lncRNAs DE et conservés entre le chien et l'homme. L'analyse $DESeq2_{multifacteur}$ étant visiblement plus stricte, les lncRNAs extraits de cette analyse peuvent correspondre à de bons candidats pour l'annotation de biomarqueurs du mélanome oral. Nous identifions ainsi 23 lncRNAs (| lFC | > 0) DE et conservés avec l'homme, ainsi que 17 lncRNAs conservés passant le seuil de | lFC | > 2 (Figure 26c. et 26d.).

3. Des lncRNAs biomarqueurs potentiels

Parmi les lncRNAs DE identifiés dans l'ensemble de l'étude ($DESeq2_{multifacteur}$ et $DESeq2_{défaut}$), nous pouvons mettre en évidence le lncRNA *ZEB2-AS1*, trouvé sur-exprimé (lFC = 2,81 et p.ajustée = $6{,}13^{e-13}$) dans les tumeurs par rapport aux contrôles (Figure 27a), ce lncRNA est déjà utilisé comme facteur pronostic du carcinome hépatocellulaire (Li *et al.*, 2016). D'après l'analyse de co-expression précédemment décrite, le lncRNA *ZEB2-AS1* et le mRNA antisens *ZEB2* sont co-exprimés (rho = 0,81 ; p.ajustée < $1{,}5^{e-4}$) et l'ensemble de ces résultats corrobore une étude de 2018 qui a montré que (i) même si le lncRNA est en antisens du gène mRNA *ZEB2*, ils pourraient partager le même promoteur et (ii) le mRNA *ZEB2* est également retrouvé impliqué dans différents cancers humains (Guo *et al.*, 2018).

RLOC_Chien	Nom_gène_Humain	ENSG_Humain		RLOC_Chien	Nom_gène_Humain	ENSG_Humain
RLOC_00020381	ENSG00000225791	TRAM2-AS1		RLOC_00014278	ENSG00000231327	AC016700.5
RLOC_00020964	ENSG00000253669	KB-1732A1.1		RLOC_00010654	ENSG00000236678	LINC00347
RLOC_00014962	ENSG00000232328	AC011286.1		RLOC_00000381	ENSG00000231772	RP1-154K9.2
RLOC_00015837	ENSG00000258602	RP11-7F17.7		RLOC_00004057	ENSG00000231842	RP11-527F13.1
RLOC_00021053	ENSG00000249917	LINC00536		RLOC_00028821	ENSG00000267299	CTB-66B24.1
RLOC_00010689	ENSG00000233379	RP11-318G21.4		RLOC_00021954	ENSG00000260049	RP3-404K8.2
RLOC_00030719	ENSG00000228420	RP4-735C1.6		RLOC_00002677	ENSG00000259172	RP11-299G20.2
RLOC_00035653	ENSG00000243885	RP11-278L15.2		RLOC_00007595	ENSG00000270344	RP11-734K2.4
RLOC_00015461	ENSG00000258946	RP11-58E21.4		RLOC_00003786	ENSG00000223598	RP1-292B18.4
RLOC_00009650	ENSG00000264443	RP4-594I10.3		RLOC_00023326	ENSG00000246695	RASSF8-AS1
RLOC_00012211	ENSG00000248714	RP11-1079K10.3		RLOC_00016972	ENSG00000260824	CTD-2235H24.2
RLOC_00008862	ENSG00000231298	LINC00704		RLOC_00001991	ENSG00000272267	RP11-375N15.2
RLOC_00001773	ENSG00000273117	AC144652.1		RLOC_00030436	ENSG00000267077	RP11-127I20.5
RLOC_00033903	ENSG00000224511	LINC00365		RLOC_00032590	ENSG00000236449	AC018890.6
RLOC_00014431	ENSG00000224511	LINC00365		RLOC_00001742	ENSG00000239911	PRKAG2-AS1
RLOC_00021932	ENSG00000231662	RP11-686D16.1		RLOC_00030709	ENSG00000224965	KCNC4-AS1
RLOC_00008826	ENSG00000215244	RP11-563J2.2		RLOC_00021002	ENSG00000253754	RP11-35G22.1
RLOC_00003079	ENSG00000250634	RP11-341G5.1		RLOC_00016055	ENSG00000258749	RP11-688G15.3
RLOC_00016575	ENSG00000259754	RP11-208K4.1		RLOC_00024295	ENSG00000258395	RP11-65I12.1
RLOC_00016200	ENSG00000258498	DIO3OS		RLOC_00003412	ENSG00000264340	RP11-231E4.3
RLOC_00010960	ENSG00000229373	LINC00452		RLOC_00026420	ENSG00000230461	PROX1-AS1
RLOC_00029422	ENSG00000273143	RP11-525A16.4		RLOC_00032620	ENSG00000224189	HOXD-AS1
RLOC_00029409	ENSG00000229981	RP11-215N21.1		RLOC_00011991	ENSG00000267042	CTD-3193K9.4
RLOC_00024263	ENSG00000172965	MIR4435-1HG		RLOC_00013021	ENSG00000224842	RP11-123K19.1
RLOC_00024210	ENSG00000234362	AC104654.2		RLOC_00031624	ENSG00000253187	HOXA10-AS
RLOC_00034983	ENSG00000233143	AC009492.1		RLOC_00033111	ENSG00000236842	RP11-399K21.10
RLOC_00022249	ENSG00000246523	RP11-736K20.6		RLOC_00032487	ENSG00000233255	AC019181.2
RLOC_00014016	ENSG00000234690	AC073283.4		RLOC_00013056	ENSG00000224020	MIR181A2HG
RLOC_00023896	ENSG00000224184	AC096559.1		RLOC_00034788	ENSG00000230408	AC007163.6
RLOC_00021785	ENSG00000223342	RP3-429O6.1		RLOC_00003622	ENSG00000227954	RP3-323P13.2
RLOC_00011772	ENSG00000263931	RP11-180P8.1		RLOC_00030536	ENSG00000260095	RP11-715J22.3
RLOC_00010227	ENSG00000223806	LINC00114		RLOC_00026355	ENSG00000227591	RP1-28O10.1
RLOC_00021953	ENSG00000272168	CASC15		RLOC_00002832	ENSG00000258384	AC068831.6
RLOC_00022942	ENSG00000250742	RP11-834C11.4		RLOC_00011768	ENSG00000265702	RP11-156L14.1
RLOC_00032366	ENSG00000228221	LINC00578		RLOC_00014809	ENSG00000235070	AC068138.1*
RLOC_00018668	ENSG00000254639	CTD-2589M5.5		RLOC_00006565	ENSG00000238290	RP11-431K24.1*
RLOC_00026042	ENSG00000203971	RP13-379L11.2		RLOC_00032685	ENSG00000260742	RP11-366L5.1*
RLOC_00006078	ENSG00000269947	RP11-849F2.9		RLOC_00020746	ENSG00000231628	RP3-355L5.4*

Table 7 : Identifiants de lncRNAs chien – humain et noms de gènes humain des 76 lncRNAs différentiellement exprimés et conservés entre le chien et l'homme. *En gras correspondent les 13 lncRNAs retrouvés en commun entre DESeq2$_{_défaut}$-DESeq2$_{(multifateur)}$-EBSeq avec | lFC | > 2. Avec *, 4 lncRNAs mis en évidence avec l'étude DESeq2$_{_défaut}$-DESeq2$_{(multifacteur)}$*

Les lncRNAs conservés identifiés comme communs entre les analyses DESeq2-EBSeq et $DESeq2_{multifacteur}$ et $DESeq2_{défaut}$) ont bien été annotés chez l'homme (Table 7) mais ils sont en revanche très peu caractérisés fonctionnellement. L'un d'entre eux, le lncRNA *CASC15* ('*Cancer Susceptibility Candidate gene 15'*), est déjà décrit comme impliqué dans le mélanome cutané humain (Lessard *et al.*, 2015). Dans notre étude, nous déterminons (Figure 27b.) que le lncRNA *CASC15* est sur-exprimé dans les tumeurs par rapport aux contrôles (lFC = 2,90 et p.ajustée = $2,86^{e-11}$). De plus, l'étude de Lessard *et al*., de 2015 a examiné le niveau d'expression de *CASC15* dans un modèle de souris de xénogreffe de métastases cérébrales issues de mélanome humain et montre que les transcrits *CASC15* sont significativement sur-exprimés (17 fois) dans les cellules de métastases cérébrales par rapport aux cellules cutanées (Lessard *et al.*, 2015), avec l'observation d'un gain de copie unique du locus *CASC15* dans les cellules métastasées. Ainsi, les auteurs concluent de ces observations que le lincRNA transcrit à partir du locus *CASC15* peut être associé à la progression métastatique du mélanome. Le *CASC15* étant conservé entre le chien et l'homme, il constitue donc un lincRNA candidat intéressant dans l'étude des mélanomes oraux canins et également humains.

Discussion - Perspectives

1. L'optimisation du dispositif expérimental

Au cours de ma thèse, nous nous sommes intéressés à la caractérisation des lncRNAs du génome canin, suite à l'identification exhaustive des lncRNAs réalisée lors du travail de Valentin Wucher (Wucher *et al.*, 2017). Nous disposons de données de RNA-seq pour 16 tissus distincts, utilisés lors de cette dernière étude, complétée par 10 autres tissus distincts. Notre dispositif expérimental est composé d'un panel relativement conséquent de 26 tissus qui représentent les principales grandes fonctions de l'organisme. Néanmoins, une augmentation du nombre de tissus permettrait d'identifier d'autres lncRNAs, qui sont pour une bonne partie tissu-spécifique, à partir de tissus encore non étudiés, comme de nouveaux tissus épithéliaux glandulaires (thyroïde, hypophyse), des tissus conjonctifs (osseux, cartilagineux, adipeux), d'autres tissus nerveux (cellules gliales, neurones). Les différents stades du développement de l'organisme seront également du plus haut intérêt à investiguer, tels que les différents stades embryonnaires, les cellules de la crête neurale (Pendleton *et al.*, 2018), les différents tissus de la spermatogénèse, le placenta, tissus dans lesquels il est probable de retrouver des lncRNAs exprimés, comme chez l'homme (Luk *et al.*, 2014; Kanduri, 2016).

Par ailleurs, nous avons réalisé une analyse de la tissu-spécificité sur des réplicats biologiques afin d'estimer la reproductibilité des données. Nous disposons de 8 échantillons de muqueuses saines issus de prélèvement buccaux de 8 Caniches. Nous avons observé lors de l'analyse de clustering que les tissus tégumentaires se regroupaient ensemble, nous avons donc analysé la tissu-spécificité uniquement sur les 20 tissus non tégumentaires, ainsi que le tissu muqueux, afin d'observer s'il existe une spécificité muqueuse en comparaison des autres tissus. Lors de cette analyse, si nous observons près de 57 % des lncRNAs muqueuse-spécifique dans plusieurs échantillons muqueux, nous avons observé que 43 % des lncRNAs muqueuse-spécifique étaient détectés dans un seul des 8 échantillons muqueux étudiés. Nous pouvons identifier ici une limite de l'étude, qui suggère que la détection de l'expression des lncRNAs n'est pas parfaitement reproductible entre réplicats biologiques, car les échantillons de tissus peuvent être hétérogènes et variables entre individus, le stade et la localisation précise du prélèvement peut varier sensiblement entre chaque

expérimentations. Pour remédier à cette limite, nous pourrions envisager différentes approches expérimentales permettant une étude plus précise et plus fine. Dans un premier temps, nous pourrions nous inspirer du projet GTEx ('Genotype-Tissue Expression') (GTEx Consortium, 2013). Ce projet permet d'étudier les variations des profils transcriptomique entre individus et tissus. Cette étude, de 2013, est réalisée sur des données issues de tissus 'non-malades'. Le consortium a séquencé par RNA-Seq 1 641 échantillons de 175 individus représentant 43 sites anatomiques : 29 tissus d'organes solides, 11 sous-régions cérébrales, le sang total et deux lignées cellulaires. Les tissus présentent une signature transcriptionelle caractéristique à la fois pour l'expression des mRNAs et des lncRNAs, les phénotypes individuels contribuant peu (Melé *et al.*, 2015). Ces signatures sont dominées par un nombre relativement faible de gènes, fait observé particulièrement dans le sang, bien que peu sont exclusives à un tissu en particulier et varient plus à travers les tissus que les individus. Cette étude a pu mettre en évidence que les gènes présentant une variation d'expression interindividuelle élevée correspondent à des candidats pathologiques associés au sexe, à l'origine ethnique et à l'âge.

Une seconde approche, appelée Single Cell, peut être utilisée dans la perspective de cette étude. Par opposition au séquençage RNA-seq en masse (bulk RNA-seq), le 'single cell RNA-seq' (scRNA-seq) est une nouvelle technologie, reportée par une première publication par Tang *et al.,* de 2009 (Tang *et al.*, 2009), et qui n'a véritablement trouvé une grande utilisation qu'à partir de 2014 quand les nouveaux protocoles et les coûts inférieurs de séquençage l'ont rendu plus accessible. Les principaux atouts du scRNA-seq sont la capacité à mesurer la distribution des niveaux d'expression pour chaque gène à travers une population de cellules. Ces données permettent d'étudier des questions pour lesquelles des modifications spécifiques au transcriptome sont importantes, telles que l'identification du type cellulaire, l'hétérogénéité des réponses cellulaires, la stochasticité de l'expression génique, ou la définition de réseaux de régulation génique à travers les cellules.

Actuellement, plusieurs protocoles différents sont utilisés, SMART-seq2 (Picelli *et al.*, 2013), CELL-seq (Hashimshony *et al.*, 2012) et Drop-seq (Macosko *et al.*, 2015). En parallèles, plusieurs instrumentations commerciales sont également disponibles, notamment Fluidigm C1, Wafergen ICELL8 et Chromium 10X Genomics. Les méthodes d'analyse computationnelle à partir du RNA–seq en masse peuvent être utilisées mais dans la plupart

des cas, l'analyse computationnelle nécessite l'adaptation des méthodes existantes ou le développement de nouvelles méthodes.

La principale différence entre le RNA-seq et scRNA-seq est que chaque banque de séquençage représente une seule cellule, au lieu d'une population de cellules. Par conséquent, une grande attention doit être accordée à la comparaison des résultats de différentes banques. Les principales sources de différences entre les banques sont que pour chaque expérimentation, un gène peut être observé à un niveau d'expression modéré dans une cellule mais n'est pas détecté dans une autre cellule (Kharchenko, Silberstein and Scadden, 2014). Dans les deux cas, les différences sont introduites en raison des faibles quantités de départ des transcrits puisque l'ARN provient d'une seule cellule. L'amélioration de l'efficacité de la capture des transcrits et la réduction du biais d'amplification sont actuellement des domaines de recherche actifs. Cependant, il est possible d'atténuer certains de ces problèmes par une normalisation et des corrections appropriées.

L'utilisation du scRNA-seq permettrait notamment de séquencer individuellement toutes les cellules composant le cycle de la spermatogénèse et d'identifier plus précisément certains phénomène lier à l'une ou l'autre des étapes. Ou encore dans le cas de tissus à forte hétérogénéité et complexité, nous pourrions différencier parfaitement les types et sous-types cellulaires qui les composent.

2. Génomique comparative et conservation des lncRNAs

Lors de notre étude sur la conservation des lncRNAs entre le chien et l'homme, nous avons pu identifier un ensemble de 939 lncRNAs conservés (~ 9 %). Cette étude permet de constituer une liste de lncRNAs potentiellement candidats lors d'étude de maladies homologues à l'homme. L'étude de la conservation des lncRNAs pourrait être envisagée entre le chien et d'autres espèces mammifères pour lesquels le répertoire de lncRNA est bien annoté (le chimpanzé, le macaque, la vache, la souris et le rat (Washietl, Kellis and Garber, 2014)). Cela pourrait, par exemple, enrichir le nombre de lncRNAs canins avec un orthologue et les connaissances sur l'évolution des lncRNAs.

Afin de compléter l'approche de synténie pour identifier les différents lncRNAs conservés, des méthodes utilisant la conservation de structure secondaire pourraient être envisagées. La conservation évolutive de la structure secondaire de l'ARN est une

caractéristique typique de nombreux ARN fonctionnels non-codants (Gruber et al., 2008). En principe, des mesures de conservation de ces structures pourraient être utiles pour les algorithmes généraux de recherche d'homologie des lncRNAs qui combinent la conservation de séquence et de structure.

Nous avons également utilisé une approche basée sur la détection des éléments contraints lors de l'analyse de conservation des lncRNAs, nommée GERP ('Genomic Evolutionary Rate Profiling') (Davydov *et al.*, 2010). L'approche GERP calcule des scores de contrainte pour des positions individuelles d'un alignement multiple, puis définit des segments correspondant à un ensemble de nucléotides contigus hautement contraints. L'absence de conservation des lncRNAs entre le chien et l'homme, peut être, en partie, expliquée par la présence d'éléments répétés (TEs) dans l'alignement multiple. Une identification et une comparaison de ces TEs dans les deux espèces permettraient de mettre en évidence plus de lncRNAs conservés entre le chien et l'homme.

Pour résumer, un ensemble d'analyses par conservation de séquence, de structure secondaire, ainsi que de bloc de synténies permettrait d'identifier un maximum de lncRNAs (et de mRNAs) conservés entre le chien et l'homme. Cette démarche d'identification permettra de compléter l'annotation du génome canin, qui progressivement répertorie de plus en plus de lncRNAs.

3. Approches complémentaires au RNA-seq

D'après l'étude de Guttman *et al*., de 2009 (Guttman *et al.*, 2009), une approche exploitant la structure de la chromatine est utilisée pour identifier les lincRNAs. Ils ont développé une méthode efficace pour créer des cartes d'état de la chromatine à l'échelle du génome, en utilisant l'immunoprécipitation de la chromatine suivie d'un séquençage massivement, nommée ChIP-Seq. Ils ont ensuite observé que les gènes activement transcrits par l'ARN polymérase II (Pol II) sont marqués par la triméthylation de la lysine 4 de l'histone H3 (H3K4me3) au niveau de leur promoteur et la triméthylation de la lysine 36 de l'histone H3 (H3K36me3) le long de la région transcrite. Ils désignent cette structure distinctive par un 'domaine K4-K36'. Ils ont ainsi proposé que, l'identification de ces structures de K4-K36 résidant à l'extérieur des loci de gènes codant pour des protéines connues, corresponde à la découverte de lincRNAs.

C’est pourquoi, nous avons réalisé une analyse à l’aide de données de ChIP-Seq issu d’échantillon de foie canin. Ces données proviennent de l’étude de Villar *et al*., de 2015 (Villar *et al.*, 2015). Cette étape de recherche de l’intersection entre nos données issues de RNA-seq et les marques de ChIP-Seq, nous a permis de croiser et de conforter les résultats de la localisation des séquences promotrices. En revanche, nos analyses sont issues d’échantillons de 26 tissus distincts, pour lesquels des marques épigénétiques par des données de ChIP-Seq pour les différents tissus conforteraient nos données.

Enfin, la technologie de CAGE (‘Cap Analysis of Gene Expression’) semble également être un bon moyen de déterminer précisément la localisation des promoteurs et des sites d’initiation de la transcription (TSS). Le but de l’approche est de cartographier de manière exhaustive la grande majorité des sites d’initiation de la transcription humaine (TSS) et donc leurs promoteurs. En comptant le nombre d'étiquettes CAGE pour chaque promoteur dans un gène, la technologie CAGE permet de déterminer également le niveau d'expression des ARN, grâce à une détection numérique de la fréquence des étiquettes. Cette approche permettrait, à la fois un contrôle de coordonnées des lncRNAs identifiés mais également des profils d’expression de ces derniers (Lizio *et al.*, 2017).

4. La co-expression lncRNA:mRNA

Lors de l’étude de la co-expression des paires lncRNA:mRNA, nous avons identifié plus de 7 600 paires lincRNA:mRNA et plus de 500 paires antisense:mRNA. Ces chiffres correspondent à plus de 3 400 lncRNAs uniques impliqués dans ces co-expressions, soit environ 30 % des lncRNAs caractérisés au total, qui cependant pourrait contenir des faux-positifs. L’analyse de co-expression entre les lncRNAs et les mRNAs est une étape clé qui permet de définir le rôle potentiel des lncRNAs en fonction des mRNAs avec lesquels ils interagissent. Pour améliorer et affiner les potentielles fonctions de ces lncRNAs, il serait intéressant de tester différents seuils de corrélation entre les expressions des lncRNAs et des mRNAs, afin de réduire le nombre de potentiels faux-positifs. Pour aller plus loin dans l’analyse de co-expression, nous avons tenté d’identifier des co-expressions en *trans*, c’est-à-dire de chercher les paires lncRNA:mRNA au delà d’1 Mb. Pour réaliser cette analyse préliminaire, nous avons analysé 2 chromosomes canins (Chr36 et Chr38) pour limiter le nombre de paires potentielles (132 et 108 lncRNAs et 177 et 230 mRNAs, pour les

chromosomes 36 et 38 respectivement). Dans un premier temps, nous avons extrait avec le module de FEELnc_classifier toutes les paires trouvées entre les chromosomes 36 et 38. Ensuite, nous avons filtré avec les mêmes critères que lors de l'analyse de co-expression dans une fenêtre de 1 Mb (| rho | > 0,5 ; p.adjust BH < 0,05), et obtenue 2 243 paires co-exprimées. Parmi ces paires, nous observons que 71,3 % des lncRNAs sont co-exprimés avec 92,8 % des mRNAs, ce qui semble élevé, et indiquent que les filtres et seuils statistiques ne sont probablement pas adaptés. Plusieurs méthodes de corrections pour tests multiples existent et devront être testées, telles que les méthodes de Benjamini-Hochberg (Cai and Chan, 2016), utilisée lors de ces analyses, de Bonferroni (Sedgwick, 2014) ou encore de Benjamini & Yekutieli (Yekutieli and Benjamini, 2001).

Suite à cette analyse, nous avons également pris en compte les co-expressions entre les lncRNAs d'un seul chromosome (Chr 38) et l'ensemble des 21 810 mRNAs. Nous obtenons 78 682 paires lncRNA:mRNA significativement co-exprimées, présentant une moyenne de 404 mRNAs co-exprimés par chromosome avec 91,7 % de lncRNAs (98 / 108). L'ampleur du nombre de paires nous a convaincu d'établir d'autres filtres, comme un seuil du niveau d'expression. Un des seuil utilisé a permis d'isoler les paires où au moins 5 gènes étaient exprimés > 1 TPM, avec cette restriction nous identifions une diminution du nombre de paires de près de 2,5 fois, passant à 28 620 paires lncRNA:mRNA. Ce chiffre est encore élevé, il serait intéressant de combiner différents filtres comme le niveau d'expression des données d'entrées, le seuil de corrélation rho ainsi que la méthode de correction pour tests multiples.

Une autre approche est de plus en plus utilisée et pourrait compléter notre analyse de co-expression. Cette approche correspond au logiciel WGCNA de R pour 'Weighted correlation network analysis'. Cette approche (disponible dans un package R) permet d'effectuer divers aspects de l'analyse de réseau de corrélation pondérée. L'ensemble comprend des fonctions statistiques pour la construction de réseaux et la détection de modules de gènes avec des patrons d'expression similaires (Langfelder and Horvath, 2008). Cette approche réalise une méta-analyse basée sur les réseaux biologiques réalisés à partir de corrélations entre deux variables.

5. Le mélanome oral canin

Suite à l'étude de la caractérisation des lncRNAs ainsi qu'à leur prédiction fonctionnelle bioinformatique, nous avons initié une étude pour identifier leurs implications potentielles dans le mélanome oral canin. Cette étude permet de proposer le chien comme modèle pour les mélanomes humains. L'analyse des données de mélanome muqueux ne présente encore que des résultats préliminaires. Néanmoins, l'étude est prometteuse grâce à un design expérimental qui rassemble 46 échantillons tumoraux issu de mélanomes oraux de trois races différentes, ainsi que de 14 échantillons appariées, c'est-à-dire que pour chacun des 14 chiens présentant un mélanome buccal, un tissu muqueux sain est prélevé.

Lors de cette analyse, deux outils ainsi que différents paramètres ont été utilisé. Dans un premier temps, nous avons comparé DESeq2 avec EBSeq. Dans nos analyses, nous avons étudié les lncRNAs au niveau du gène. Or EBSeq est fréquemment utilisé lors de l'étude de différentiel d'expression avec les données d'isoformes, même s'il est démontré comme robuste avec des données de gènes (Seyednasrollah, Laiho and Elo, 2015). Par la suite, il serait intéressant d'effectuer les analyses de cette étude au niveau des isoformes des lncRNAs canins.

Dans un second temps, l'analyse s'est focalisée sur deux effets pouvant intervenir dans le différentiel d'expression, il s'agit de l'effet race ainsi que de l'effet type cellulaire. Comme nous avons pu le démontrer précédemment, l'approche par Single Cell serait très intéressante dans notre étude. Dans ce cas, le fait de séquencer le type cellulaire précisément, nous pourrions nous affranchir de la condition présentant les deux types cellulaires kératinocyte et mélanocyte.

Pour la poursuite de l'étude des mélanomes buccaux canins, l'intégration de données précliniques canines ainsi qu'humaines nous permettrait d'approfondir nos connaissances sur les lncRNAs potentiellement biomarqueurs. La recherche d'ADN circulant ou encore appelé les cellules cancéreuses circulantes (CTC, pour 'Circulating Tumours Cells'), serait également un bon moyen de détecter des lncRNAs connus dans l'implication de certains cancers, afin de les utiliser en tant que facteurs diagnostics ou pronostics de cancers. Dans l'étude de Hegemann *et al.*, de 2016 (Hegemann *et al.*, 2016), cette approche a été approuvée comme test de diagnostic dans le cas du cancer de la prostate. Le nombre de CTC

s'est avéré pertinent pour le pronostic du cancer de la prostate au stade métastatique (Kretschmer and Tilki, 2017).

Comme exposée dans l'introduction, l'histoire du chien de par son évolution et les pratiques de sélection des races, en fait un modèle unique. La forte homogénéité génétique au sein des races a entrainé l'enrichissement de nombreux variants, spécifiques à chaque race. Ces caractéristiques permettent aujourd'hui d'identifier les bases de maladies génétiques. Parmi les altérations génétiques récurrentes chez l'homme, certaines se sont révélées être informatives sur le diagnostic et le pronostic. En plus de cela, elles peuvent présenter un réel potentiel thérapeutique. Beaucoup de molécules sont disponibles aujourd'hui et décrites comme inactivant les oncogènes dans différentes voies oncogéniques ou restaurant l'activité de suppresseurs de tumeur. Cependant, dans la pratique, les innovations thérapeutiques peinent à suivre, lorsque ces molécules ont prouvé leur activité en préclinique et présentent peu de risque toxicologique, il arrive souvent qu'au stade clinique, la grande majorité échoue du fait d'un manque d'efficacité, ce qui constitue une véritable limite. Cependant, il est désormais possible de développer des essais précliniques chez le chien. Même si l'inclusion de chiens de compagnie existe dans le développement de molécules thérapeutiques pour la médecine humaine et vétérinaire aux Etats-Unis, elle reste encore rare en France. Compte tenu du nombre de chiens de compagnie en France, de la médecine vétérinaire poussée avec l'existence de centres spécialisés dans le traitement et le suivi des cancers canins, la communauté scientifique ne devrait pas ignorer cette opportunité déjà largement saisie aux Etats-Unis. De plus, un des grands avantages du modèle chien, est que les propriétaires de chien peuvent réaliser le traitement à la maison et leurs chiens peuvent être ensuite suivis par des vétérinaires. Le chien étant, après l'homme, le plus suivi au niveau médical, une étude allant du diagnostic à la rémission en passant par un suivi post-traitement, voir l'observation de rechute, peut également être réalisée sans grandes difficultés. Ces études pourront également profiter aux patients humains dans une logique de médecine transversale.

Qui plus est, les mêmes gènes peuvent être altérés avec des fréquences différentes entre le chien et l'homme. Ainsi, le chien peut devenir un modèle pertinent pour des sous-types particuliers. Cela peut être illustré par le gène *NRAS* qui est muté dans les mélanomes buccaux, mutations qui correspondent aux mutations retrouvées chez l'homme

(Gillard *et al.*, 2014). Par conséquent, les mélanomes canins représentent des modèles nouveaux et pertinents pour des sous-groupes de mélanomes chez l'homme. Les altérations et les récurrences des mélanomes canins en font un modèle prometteur pour le développement de médicaments et la mise en place d'essais cliniques vétérinaires.

6. La validation fonctionnelle

Tout le travail effectué lors de cette thèse constitue un ensemble de données sur la caractérisation des lncRNAs et les prédictions fonctionnelles bioinformatiques. Cette étude reste, cependant, à un niveau '*in silico*'. Pour aller plus loin, une validation fonctionnelle de certains de ces lncRNAs serait très informative, notamment pour les lncRNAs différentiellement exprimés entre la tumeur et le contrôle que nous retrouvons également conservés avec l'homme. À cette étape de l'analyse, nous présentons des prédictions, des suggestions mais pas d'affirmations avec preuve expérimentales. Pour certains mécanismes identifiés, notamment lorsqu'il s'agit de co-expression d'une paire lncRNA:mRNA, nous pourrions réaliser l'inactivation du lncRNA co-exprimé afin d'observer l'expression du mRNA voisin.

Conclusion

Au cours de ces trois années de thèse, nous nous sommes intéressés à l'analyse intégrative des lncRNAs en étudiant leurs profils d'expression, leurs contenus en éléments transposables, leurs conservations avec l'homme, leurs co-expressions avec les mRNAs, ainsi que leurs implications potentielles dans le mélanome oral. Nous reportons dans cette étude l'analyse d'expression de plus de 10 000 lncRNAs canins dans 26 tissus distincts. Les résultats montrent que les lncRNAs canins, comme pour les autres espèces, sont exprimés plus faiblement que les mRNAs mais de manière spécifique à un tissu. Ainsi, nous avons identifié ici les signatures transcriptionnelles spécifiques aux 26 tissus analysés. La forte spécificité tissulaire des lncRNAs peut constituer un facteur limitant pour annoter le répertoire des lncRNAs d'une espèce puisque leur annotation va ainsi dépendre de la disponibilité de nombreux types cellulaires, de l'analyse de nombreux tissus et de multiples conditions de temporalité, pour définir de manière exhaustive le catalogue complet des lncRNAs et de leur isoformes.

L'étude des réseaux de co-expression révèle plus de 3 400 lncRNAs uniques co-exprimés avec les mRNAs proches et met en évidence de possibles processus de *cis*-régulation géniques. Par une approche de génomique comparative, nous identifions plus de 900 lncRNAs avec un orthologue humain et l'analyse du contenu en éléments transposables montre que près de 70 % des exons des lncRNAs canins chevauchent au moins un TE. Ces observations suggèrent que les lncRNAs canins évoluent rapidement en terme de séquence (insertion de SINEC_Cf) et ainsi se différencient pour devenir des éléments fonctionnels spécifiques du développement et de la régulation de fonctions cellulaires spécialisées.

Enfin les résultats préliminaires sur l'étude des lncRNAs impliqués dans le mélanome oral canin, nous ont permis d'identifier un ensemble de lncRNAs différentiellement exprimés entre la tumeur et le contrôle. Nous mettons en évidence plus de 750 lncRNAs différentiellement exprimés entre la tumeur et le contrôle (FDR < 0,01), ainsi que plus de 100 lncRNAs parmi ces derniers présentant une conservation entre le chien et l'homme.

Cette étude indique que les lncRNAs appartiennent à de multiples classes fonctionnelles et suggère des fonctions potentielles aux lncRNAs, dans des processus

fondamentaux tels que la spermatogenèse et le développement, ou encore dans des mécanismes plus spécifiques tels que le développement d'organes sensoriels et la croissance cellulaire. Enfin, ces lncRNAs présentent également de potentielles implications dans la progression tumorale du mélanome oral, homologue aux mélanomes muqueux chez l'homme, suggérant une transversalité des connaissances et des applications de potentiels biomarqueurs.

L'ensemble de cette étude apporte une meilleure annotation du génome canin et une caractérisation exhaustive des profils d'expression des ARN longs non-codants, ainsi qu'une première mise en évidence de leurs potentiels rôles fonctionnels par une approche de bioinformatique.

Matériels et Méthodes

I. Identification et caractérisation des lncRNAs canins

1. Données initiales issues de 26 RNA-seq

Mon travail de thèse est basé sur des données de séquençages complets de transcriptome (RNA-seq). Tout d'abord, mon projet concernant la caractérisation et la prédiction fonctionnelle bioinformatique des lncRNAs porte sur 26 échantillons canins de systèmes anatomiques distincts, répertorié dans le tableau suivant (Table 3), permettant d'annoter et de caractériser les gènes canins.

Dans un premier temps, les prélèvements ont été obtenus via la biobanque française « CRB Cani-DNA[10] » de l'Université de Rennes1, crée par Catherine André au sein de l'équipe, la biobanque de l'Université de Copenhague, au Danemark, la biobanque de l'Université d'Helsinki en Finlande et la biobanque Vetsuisse de l'Université de Bern, en Suisse. Les ARN ont été extraits des tissus en utilisant le kit NucleoSpin RNA (Macherey-Nagel) selon les instructions du fabricant. Des ARN polyadénylés ont été capturés par des billes d'oligodT et des librairies de RNA-seq ont été construites par l'intermédiaire du kit de préparation d'échantillons de mRNA à brins orientés, TruSeq™ d'Illumina. Le séquençage a été effectué de manière appariée et orientée sur un instrument de type HiSeq2000 en utilisant la chimie v3 (TruSeq PE Cluster Kit v3-cBot-HS, TruSeq SBS Kit v3-HS, TruSeq SR Cluster Kit v3-cBot-HS) à une profondeur d'environ 50 millions de reads par tissu.

Les 26 tissus sains sont regroupés par systèmes anatomiques selon la classification proposée par la base de données Expression Atlas[11]. Au total, deux tissus (intestin_côlon et jéjunum) sont dans le système digestif, six (follicule pileux, glandes mammaires, kératinocyte, muqueuse, peau et planum nasal) dans le système tégumentaire, deux (muscle squelettique, muscle cardiaque) dans le système musculo-squelettique et cinq (bulbe olfactif, cerveau, cervelet, cortex et moelle épinière) dans le système nerveux. Les autres

[10] https://dog-genetics.genouest.org/cani-dna/
[11] https://www.ebi.ac.uk/gxa/home/

tissus (foie, glandes surrénales, pancréas, poumons, rate, rein, rétine, sang et tissue ovarien) correspondent à la catégorie 'Unmapped' et au tissu testiculaire.

2. Description de l'annotation de référence du génome canin (CanFam3.1-plus)

Basé sur l'annotation étendue de CanFam3.1-plus, nous travaillons sur un jeu de données comprenant au total 10 444 lncRNAs et 21 810 mRNAs, correspondant respectivement à 22 880 et 158 750 transcrits. Ces deux biotypes, non-codant et codant ont été identifiés dans une précédente étude avec l'outil FEELnc (Wucher *et al.*, 2017). Dans cette étude, réalisée au sein de l'équipe, il a été identifié 2 436 nouveaux lncRNAs et 10 374 nouveaux transcrits ainsi que 336 nouveaux mRNAs et 58 640 transcrits (Table 1). Elle a permis d'améliorer de manière générale l'annotation précédente (CanFam3.1), en doublant le nombre de transcrits lncRNAs et en augmentant le nombre de transcrits des mRNAs de 50 %. Cette étude a également permis de découvrir que les transcrits étaient plus longs en termes du nombre d'exons, de CDS / UTRs et de tailles des ARN, suggérant une reconstitution plus complète des structures de gènes.

3. Intersection de l'annotation Canfam3.1-plus et des promoteurs et enhancers

Une des premières étapes que j'ai réalisée au cours de ma thèse, a consisté en la comparaison de positions des enhancers et promoteurs détectés par des analyses de ChIP-seq générés à partir d'échantillons du tissu foie avec l'annotation du génome canin. Au cours de mon stage de Master 2 à l'EBI ('European Bioinformatic Institute') à Hinxton, j'ai étudié les enhancers et promoteurs chez 5 mammifères, dont le chien dont j'ai pu analyser les données. J'ai utilisé les coordonnées génomiques de ces éléments disponibles via l'étude de Villar *et al*., (Villar et al. 2015). Cette étude rapporte que l'évolution rapide des enhancers est une caractéristique universelle des génomes de mammifères et révèlent qu'au contraire les promoteurs sont partiellement ou entièrement conservés dans les 20 espèces étudiées. Pour cette analyse, j'ai utilisé l'annotation du génome canin CanFam3.1-plus sous la forme d'un fichier `.gtf` (Wucher *et al.*, 2017) et deux fichiers au format `.bed` avec les coordonnées de 30 679 marques de modifications des Histones 3, la marque d'acétylation de la lysine 27 (H3K27ac) pour les enhancers et 11 099 marques de triméthylation de la lysine 3 (H3K4me3) pour les promoteurs (Villar *et al.*, 2015). Plusieurs intervalles de séquences promotrices ont été considérés allant de 1, 5 ou 10 kb en 5' de chacun des sites

d'initiation de la transcription (TSS) des gènes. Ensuite en utilisant l'outil bedtools intersect de la suite BEDTools (Quinlan and Hall, 2010), j'ai réalisé l'intersection à un nucléotide près de l'annotation du génome canin et des marques H3K27ac et H3K4me3, à l'aide de la commande suivante :

```
bedtools intersect -a $enhancerORpromoter.bed -b $canineGenome.gtf -wa -wb -f
1E-9 > output_file
```

où –a et –b correspondent aux fichiers d'entrées où chaque entité de A est comparée avec chaque entité de B et `-wa` et `-wb` correspondent à l'option qui permet de regarder le chevauchement de A sur B (`-wa`) et de B sur A (`-wb`). Afin de valider le chevauchement des marques d'histones spécifiques aux promoteurs (H3K4me3) et aux enhancers (H3K27ac), nous avons générés aléatoirement des marques de promoteurs et d'enhancers. Grâce à la commande suivante, nous avons pu créer 100 jeux de données de marques ChIP-seq aux coordonnées aléatoires sur le génome canin :

```
bedtools shuffle -i $FILE -g $canFam3.chr.sizes -noOverlapping > $output
```

4. Pipeline d'analyse bioinformatique des données de RNA-seq

L'analyse des données de séquences RNA-seq est réalisée selon les étapes bioinformatiques décrites dans Dejbali *et al.,* 2017 (Djebali *et al.*, 2017) (Figure 1). L'alignement des reads ('mapping') a été réalisé par le programme STAR (Dobin *et al.*, 2013) (version 2.5.0a), lequel nécessite la création d'un index du génome et du transcriptome afin d'accélérer ce processus. Puis, j'ai utilisé STAR avec les options suivantes :

```
STAR --genomeDir $STARgenomeDir --readFilesIn $read1 $read2 --
readFilesCommand zcat --outFilterType BySJout --outSAMunmapped Within --
outSAMtype BAM SortedByCoordinate --limitBAMsortRAM 34146942384 --
outSAMattrIHstart 0 --outFilterIntronMotifs RemoveNoncanonical --
alignEndsType EndToEnd --runThreadN $nThreadsSTAR --quantMode
TranscriptomeSAM --outWigType bedGraph --outWigStrand Stranded --outTmpDir
$tmpDir --outFileNamePrefix $jobName
```

où `$nThreadsSTAR` = 4 est le nombre CPUs utilisés pour parralléliser le mapping. Cette étape a été réalisée pour les 26 données de RNA-seq aboutissant à 26 fichiers au format `.bam`.

Pour finir, j'ai réalisé le comptage des reads sur l'annotation CanFam3.1-plus à l'aide de l'outil RSEM (Li and Dewey, 2011) (version 1.2.25). Un index du génome est réalisé, puis le comptage des reads est effectué avec les options suivantes :

```
rsem-calculate-expression --bam --no-bam-output --estimate-rspd --calc-ci --
seed 12345 -p $nThreadsRSEM --ci-memory 30000 --paired-end --forward-prob 0
$input $RSEMgenomeDir $outName
```

Cette analyse aboutit à un fichier comprenant 17 colonnes dont une correspondant au niveau d'expression normalisé de chaque gène en TPM ('Transcript Per Millions'). Une concaténation de ces 26 fichiers de comptage des reads en TPM est réalisé afin d'obtenir une matrice d'expression pour les 10 444 lncRNAs et 21 810 mRNAs.

5. Calcul de la tissu-spécificité d'un gène

Afin d'étudier la tissu-spécificité des gènes lncRNA et mRNAs, j'ai calculé un score de spécificité nommé 'tau' (Yanai *et al.*, 2005), compris entre 0 et 1, où 1 correspond lorsque le gène est fortement tissu-spécifique. En utilisant la matrice d'expression décrite précédemment, le calcul de tau est réalisé selon l'équation suivante :

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

où n correspond au nombre de tissus analysé et x_i l'expression du gène x dans le tissu i. Pour cette analyse, nous avons considéré uniquement les gènes ayant un niveau d'expression minimum > 1 TPM dans au moins un tissu afin d'éliminer les gènes faiblement exprimés. Un script perl m'a permis de calculer l'indice tau pour l'ensemble des gènes de la matrice d'expression normalisée en TPM de la façon suivante, disponible sur le site GitHub[12] :

```
perl specificity_score.pl file.tpm 1 > outputFile.tpm
```

Pour extraire les gènes les plus tissu-spécifiques, le seuil de 0,95 a été fixé et correspond, en moyenne, à un ratio de 4 entre le plus haut niveau d'expression du gène x dans un tissu i et le deuxième plus haut niveau d'expression du même gène x dans un tissu i', où en d'autres termes :

$$\frac{1^{er}\max(xi)}{2^{nd}\max(xi')} > 4$$

[12] https://github.com/tderrien/IGDR/blob/master/script/specificity_score.pl

6. Clustering ascendant hiérarchique

Pour réaliser le clustering ascendant hiérarchique des données d'expression des lncRNAs et mRNAs séparément, j'ai calculé les corrélations de Spearman par paires entre les 26 tissus afin d'obtenir une matrice de distance. Grâce à cette dernière, la heatmap ainsi que le dendrogramme associé sont créés. Cette étape utilise la fonction R `hclust` avec les paramètres par défaut : une distance euclidienne et la méthode d'agrégation dite '`complete`'. Un fichier complémentaire indique pour chaque tissu le système anatomique dans lequel il est classé (système digestif, tégumentaire, musculo-squelettique, nerveux, 'unmapped' et testiculaire). Le script permettant de réaliser la figure 15 est inspiré du modèle décrit dans Breschi *et al.*, 2016 (Breschi *et al.*, 2016) utilisant le logiciel R (version 3.2) et est disponible via le site GitHub[13].

7. Analyse du contenu en Éléments Transposables (TEs)

Les annotations des TEs du génome canin sont basées sur le programme RepeatMasker (Smit, AFA, Hubley, R & Green, 2013) et ont été téléchargées à partir de la base de données UCSC (Casper *et al.*, 2018) (Septembre 2017). Quatre classes principales ont été analysées : les transposons à ADN et les rétrotransposons de type 'long interspersed elements' (LINEs), 'long terminal repeats' (LTRs) et 'short interspersed elements' (SINEs) ; chacun d'eux étant divisé en familles (par exemple SINEC_Cf pour les SINEs spécifiques des carnivores). Pour calculer la proportion de TEs par gène, les exons de chaque gène ont été projetés sur le génome canin et la proportion des nucléotides des exon-projetés chevauchant les familles de TEs a été calculée en utilisant le programme bedtools (version 2.19.0) sur les fichiers d'entrée `bed12` avec les options suivantes : `-wao` et `-split`. Pour les promoteurs de gènes, le même processus a été réalisé en utilisant 5 kb en 5' de chacun des gènes.

8. Le contenu en TE et l'expression des gènes

L'outil mixOmics (Rohart *et al.*, 2017) (version 6.3.1) a été utilisé pour calculer les corrélations canoniques et mettre en évidence les corrélations potentielles entre deux matrices *X* et *Y*, où *X* est la matrice d'expression des lncRNAs (ou mRNAs) dans les 26 tissus

[13] https://github.com/abreschi/Rscripts/blob/master/ggheatmap.R

et *Y* est la matrice comprenant le taux en TEs dans les exons et promoteurs des lncRNAs (ou mRNAs). La fonction `rcc` du package mixOmics (avec method = '`ridge`' et `ncomp` = 3) a été utilisée pour rechercher des combinaisons linéaires des variables (variables canoniques) afin de réduire la dimension des ensembles de données. La heatmap a été construite avec la fonction mixOmics `cim` ('Clustered Image Maps') avec des valeurs par défaut, exceptée `ncomp` = 3.

9. Génomique comparative et conservation évolutive des lncRNAs / mRNAs canins

9.1. Approche par synténie

Pour identifier les lncRNAs et les mRNAs canins conservés avec l'homme, j'ai utilisé la base de données EnsEMBL Compara (Herrero *et al.*, 2016). Le programme annote des régions orthologues entre deux espèces ou plus sur la base d'alignements multiples de génomes entiers ('Whole Genome Alignment'– WGA) calculés en utilisant le pipeline EPO (Enredo-Pecan-Ortheus) (Herrero *et al.*, 2016). Toutes les coordonnées des gènes canins ont ensuite été transposées sur le génome humain (version d'assemblage GRCh37 et version d'annotation EnsEMBL 75, GENCODE version 19). Cette étape permet de retrouver les différents blocs de synténie présents entre le chien et l'homme. Les gènes orthologues retenus sont ceux présentant une relation de type 1:1 pour un même biotype (codant ou non-codant), c'est-à-dire que pour un gène canin, un seul gène humain est retrouvé en synténie.

9.2. Par utilisation du score GERP

L'API EnsEMBL Compara ('Application Program Interface') utilisé à cette étape, sert d'intermédiaire entre la base de données MySQL sous-jacente et le script de l'utilisateur. Elle vise à fournir un accès de haut niveau à la base de données. Cette dernière est multi-espèces et stocke les résultats des comparaisons d'espèces à l'échelle du génome calculées pour chaque version de données à différents niveaux : (i) au niveau séquence d'ADN avec l'alignements du génome entier, les régions de synténie et les scores de conservation et éléments contraints ; (ii) au niveau gène avec des arbres phylogénétiques et des prédictions d'homologie. L'API EnsEMBL Compara permet donc de récupérer les scores GERP (Davydov *et al.*, 2010) de l'assemblage du génome du chien (CanFam3) (un fichier `.wig` pour chaque

chromosome du chien) calculé sur les 53 mammifères euthériens. Pour obtenir une seule valeur par gène lncRNA et mRNA, tous les exons de chaque gène ont été projetés sur le génome du chien et le score médian de GERP a été calculé pour chacun des gènes.

10. Transcriptomique comparative et conservation des profils d'expression chien / homme

Pour réaliser cette analyse, j'ai extrait les données d'expression des gènes humain à partir des données produites par le projet ENCODE (ENCODE Project Consortium, 2012). Sur le site ENCODE[14], j'ai utilisé les filtres suivants : (i) Organism – Homo sapiens ; (ii) Biosample type – tissue ; (iii) Assay – total RNA-seq ; (iv) Genome assembly (visualization) – hg19 and (v) Project - ENCODE. Ensuite j'ai comparé les tissus humains disponibles avec ceux utilisés chez le chien dans notre étude afin qu'ils soient communs aux deux espèces. Onze tissus communs ont ainsi été sélectionnés : la glande surrénale, l'intestin-côlon, le cœur, le foie, le poumon, le muscle, le tissu ovarien, le pancréas, la peau, la rate et le tissu testiculaire. À partir des fichiers RSEM disponibles sur le site d'ENCODE, les niveaux d'expression des transcrits normalisés en TPM ont été extraits de la même manière que pour les données du chien. Dans le cas de réplicas tissulaires humaines, c'est-à-dire lorsque plusieurs fichiers RSEM sont disponibles pour un seul tissu, la moyenne du niveau d'expression par gène a été réalisée. J'ai ensuite calculé la corrélation de Spearman, du niveau et de la distribution d'expression des lncRNAs orthologues entre les onze tissus canins et humains. Suite à ce test de corrélation, des diagrammes à barres ont été créés avec le logiciel R pour aider à visualiser cette comparaison.

11. Représentation circulaire des données de transcriptomique comparative avec Circos

L'outil Circos (circus-0.69-6) (Krzywinski *et al.*, 2009) a été utilisé pour visualiser les relations entre les gènes orthologues entre le chien et l'homme. Des schémas circulaires ont été construits pour permettre une représentation globale avec une figure Circos pour chaque chromosome canin. Pour chacun des 38 autosomes et le chromosome X, une figure Circos est réalisé avec, à droite, les relations d'orthologies chien-homme des lncRNAs, leurs

[14] https://www.encodeproject.org/matrix/?type=Experiment

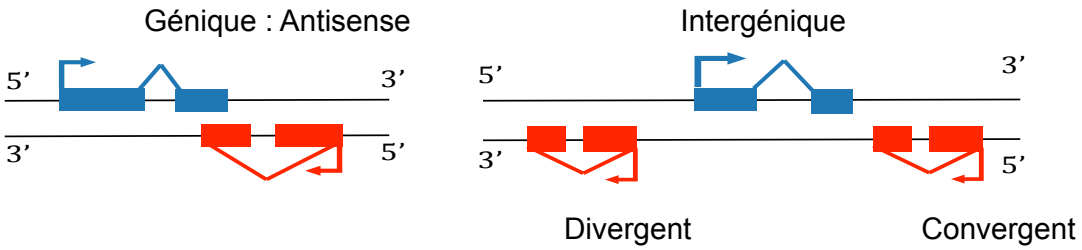

Figure 28 : Schémas représentant les antisenses (à gauche) et les lincRNAs (à droite), *différenciant les divergents des convergents.*

niveaux d'expression, puis à gauche, se trouve la représentation des 23 chromosomes humains. Les différentes couches d'informations Circos sont composées, de l'extérieur vers l'intérieur, de la manière suivante (Cf. Figure 19) :

Couche 1 : les identifiants des lncRNAs canins orthologues des lncRNAs humains.

Couche 2 : le niveau d'expression des lncRNAs canins indiqué par l'histogramme bleu.

Couche 3 : le tissu dans lequel le lncRNA est exprimé chez le chien. Onze tissus comprenant la glande surrénale, l'intestin-côlon, le cœur, le foie, le poumon, le muscle, le tissu ovarien, le pancréas, la peau, la rate et tissu testiculaire.

Couche 4 : l'expression des lncRNAs orthologue humain. Onze lignes avec un carré allant du rouge foncé lorsque l'orthologue lncRNA humain est exprimé dans le même tissu et avec la plus haute expression, au carré gris lorsque le lncRNA humain orthologue n'est pas exprimé dans le même tissu.

Couche 5 : l'arc de cercle vert représente le chromosome canin et le rouge représente les chromosomes humains.

Couche 6 : les lignes colorées relient les relations orthologues chien-homme des lncRNAs et les lignes grises représentent les relations orthologues des mRNAs.

12. Analyse de la co-expression des paires lncRNA:mRNAs canins

Pour étudier les co-expressions entre les lncRNAs et les mRNAs canins, j'ai utilisé le module de classification de FEELnc : $FEELnc_{classifier}$ (Wucher *et al.*, 2017) afin d'annoter les différentes classes de lncRNAs (comme par exemple génique versus intergénique et divergent versus convergent, Figure 28). $FEELnc_{classifier}$ permet ainsi d'identifier automatiquement tous les paires lncRNA:mRNA dans une fenêtre glissante de 1 Mb autour de chaque lncRNA. La commande utilisée est la suivante :

```
FEELnc_classifier.pl -i lncRNA_exons.gtf -a mRNA_exons.gtf -w 1000000 -m 1000
000 > lncRNA-VS-mRNA_classes.txt
```

Pour faciliter l'interprétation des résultats, les classes ont été filtrées de manière à ne conserver que les lncRNAs de type 'génique antisense exonique', 'intergénique divergent' et 'intergénique convergent'. Les 'intergénique same strand' sont, comme leur nom l'indique, orientés dans le même sens que les mRNAs, et ne sont pas pris en compte pour éviter de

comptabiliser des paires où le lncRNA aurait été mal annoté, c'est-à-dire que les exons de ce dernier seraient en réalité des exons du mRNA voisin. Pour les 126 000 paires lncRNA:mRNA obtenues avec FEELnc$_{classifier}$, j'ai calculé les corrélations de Spearman entre les vecteurs d'expression des lncRNAs et des mRNAs dans les 26 tissus avec la fonction `cor.test` de R. Les p-values des tests ont été corrigés pour les tests multiples en utilisant la méthode de Benjamini-Hochberg (BH) (Cai and Chan, 2016). Pour identifier les paires significativement co-exprimées, j'ai fixé un coefficient de corrélation à | rho | > 0,5 et une p-value ajustée p.ajustée < 0,05.

13. Analyse des enrichissement en termes GO ('Gene Ontology')

À partir des 5 711 mRNAs canins qui ont été identifiés comme significativement co-exprimés avec leurs lncRNAs voisins (n = 3 410), j'ai extrait les identifiants de gènes humains orthologues des mRNAs (n = 3 977) à partir de la base EnsEMBL. Pour identifier les termes de l'ontologie génique (GO) significativement enrichis dans cette liste d'identifiants de gènes, j'ai utilisé le serveur WebGestalt ('GEne SeT AnaLysis Toolkit' (Wang *et al.*, 2017)) avec les paramètres suivants : *Method of Interest = 'Overrepresentation Enrichment Analysis' (ORA)* et *Functional Database class = geneontology (GO) named 'Biological_Process_noRedundant'*. Enfin, j'ai sélectionné le type d'identifiant de gène *'ensemble_gene_id',* pour télécharger les données orthologues en utilisant les identifiants ENSG et j'ai sélectionné *'genome'* pour *Reference Set for Enrichment Analysis*. J'ai ajouté des paramètres avancés, dont j'ai gardé les paramètres par défaut (*Minimum Number of Genes for a Category: 5* ; *Maximum Number of Genes for a Category: 2000* et *Multiple Test Adjustment: BH*) et j'ai modifié *the Significant Level: FDR 0.05*.

II. Implication des lncRNAs dans le mélanome oral canin

1. Données initiales issues de 60 RNA-seq

La dernière partie de ma thèse a consisté à déterminer l'implication des lncRNAs, précédemment caractérisés, dans le mélanome oral canin. Cette étude a été menée grâce à l'analyse de 60 RNA-seq issus de 46 échantillons tumoraux buccaux canins, ainsi que 14 prélèvements de tissus sains. Les prélèvements tumoraux ont été réalisés sur 10 Caniches, 17 Labradors et 19 Golden Retrievers. Concernant les 14 contrôles, ils

correspondent à du tissu buccal présumé sain sur des chiens atteints de mélanomes oraux, les prélèvements sont issus des mêmes 10 Caniches et de 4 des 19 Golden Retrievers, permettant un design expérimental de données appariées. Ces prélèvements ont été obtenus via la biobanque française 'CRB Cani-DNA'.

De la même manière que pour les échantillons précédents, les ARN ont ensuite été extraits des tissus en utilisant le kit NucleoSpin RNA (Macherey-Nagel) selon les instructions du fabricant. Des ARN polyadénylés ont été capturés par des billes d'oligodT et des librairies de RNA-seq ont été construites par l'intermédiaire du kit de préparation d'échantillons de mRNA à brins orientés, TruSeq™ d'Illumina. Le séquençage a été effectué de manière appariée et orientée sur la plate-forme HiSeq2000 en utilisant la chimie v3 (TruSeq PE Cluster Kit v3-cBot-HS, TruSeq SBS Kit v3-HS, TruSeq SR Cluster Kit v3-cBot-HS) à une profondeur d'environ 50 millions de reads par tissu.

2. Recherche de lncRNAs différentiellement exprimés

L'analyse des données de séquences RNA-seq a été réalisée de la même manière que pour les 26 tissus sains de l'étude précédente. L'alignement des reads ('mapping') a été réalisé par le programme STAR (Dobin *et al.*, 2013) (version 2.5.0a) et le comptage des reads sur l'annotation CanFam3.1-plus à l'aide de l'outil RSEM (Li and Dewey, 2011) (version 1.2.25). Une analyse en composante principale (ACP) est réalisée à partir des données d'expression pour visualiser la répartition des échantillons entre tumeurs et contrôles. Cette ACP est créée à l'aide de la fonction `plotPCA` de R. Afin de réaliser l'analyse de DE sur notre jeu de données, nous avons combiné deux outils ayant une approche légèrement différente, l'outil EBSeq (Leng *et al.*, 2013) et l'outil DESeq2 (Love, Huber and Anders, 2014).

2.1. L'outil DESeq2

Les données de comptage brutes issues de RSEM ('expected counts') des lncRNAs sont intégrées à l'outil DESeq2 qui réalise une normalisation ('estimateSizeFactors') inter-échantillons. Pour le design expérimental, les valeurs de comptage (brutes et normalisées) et l'équation ('design formula') sont stockées dans un objet R appelé 'DESeqDataSet'. Par défaut, DESeq2 va tester un seul facteur (*e.g.* la condition tumorale versus contrôle dans le cadre de l'analyse des cancers) qui explique les valeurs de comptage

du design expérimental. Le différentiel d'expression cherche à expliquer le statut (tumeur/contrôle). Cependant, plusieurs facteurs peuvent influencer les valeurs de comptage (*e.g.* effet batch, sexe, âge…). Dans l'analyse DESeq2$_{multifacteur}$, nous avons pris en compte deux autres facteurs pouvant contribuer aux valeurs de comptage : effet race (3 races présentes dans notre étude : Labrador, Golden Retriever et Caniche) et l'effet population cellulaire (kératinocytes versus mélanocytes). Pour ce dernier, nous avons utilisé le programme xCell (Aran, Hu and Butte, 2017) qui calcule l'enrichissement en différents types cellulaires (incluant kératinocytes et mélanocytes) en se basant sur une méthode de signatures de gènes entraînée sur des données issues d'expériences Single Cell RNA-seq. La commande ci-dessous permet de combiner les 3 facteurs utilisés dans notre mode DESeq2$_{multifacteur}$:

```
formula(~cell_ratio + breed + condition)
```

à la suite, un FDR (ou p-value ajustée) < 0,01 est également appliqué afin de comparer les lncRNAs identifiés par EBSeq.

2.2. L'outil EBSeq

Pour le second outil, les données de comptage des lncRNAs sont également utilisées, en revanche l'outil ne contrôle pas pour l'effet race (3 races présentes), ni pour l'effet de type cellulaire. La commande utilisée pour lancer le programme est la suivante :

```
EBSeq_output <- EBTest(Data=Matrix, NgVector=geneName,
Conditions=factor(rep(c("T","H"), each=(ncol(Matrix)/2)), levels=c("T","H")),
sizeFactors= MedianNorm(Matrix), maxround=9)
```

où 'T' et 'H' correspondent aux conditions 'Tumor' (tumeur) et 'Healthy' (contrôle). Par la suite un FDR ('False Discovery Rate') = 0,01 est appliqué.

Bibliographie

1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, et al. 2015. 'A Global Reference for Human Genetic Variation.' *Nature* 526 (7571): 68–74. doi:10.1038/nature15393.

Abadie, Jérôme, Benoit Hédan, Edouard Cadieu, Clotilde De Brito, Patrick Devauchelle, Catherine Bourgain, Heidi G Parker, et al. 2009. 'Epidemiology, Pathology, and Genetics of Histiocytic Sarcoma in the Bernese Mountain Dog Breed.' *The Journal of Heredity* 100 Suppl (Suppl 1). Oxford University Press: S19-27. doi:10.1093/jhered/esp039.

Adams, V. J., K. M. Evans, J. Sampson, and J. L. N. Wood. 2010. 'Methods and Mortality Results of a Health Survey of Purebred Dogs in the UK.' *The Journal of Small Animal Practice* 51 (10). Wiley/Blackwell (10.1111): 512–24. doi:10.1111/j.1748-5827.2010.00974.x.

Ahrendt, S A, S Halachmi, J T Chow, L Wu, N Halachmi, S C Yang, S Wehage, J Jen, and D Sidransky. 1999. 'Rapid P53 Sequence Analysis in Primary Lung Cancer Using an Oligonucleotide Probe Array.' *Proceedings of the National Academy of Sciences of the United States of America* 96 (13). National Academy of Sciences: 7382–87.

Al-Tobasei, Rafet, Bam Paneru, and Mohamed Salem. 2016. 'Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout.' *PloS One* 11 (2): e0148940. doi:10.1371/journal.pone.0148940.

Alnasir, Jamie, and Hugh P Shanahan. 2015. 'Investigation into the Annotation of Protocol Sequencing Steps in the Sequence Read Archive.' *GigaScience* 4. Oxford University Press: 23. doi:10.1186/s13742-015-0064-7.

Alzohairy, Ahmed M., Gábor Gyulai, Robert K. Jansen, and Ahmed Bahieldin. 2013. 'Transposable Elements Domesticated and Neofunctionalized by Eukaryotic Genomes.' *Plasmid* 69 (1). Academic Press: 1–15. doi:10.1016/j.plasmid.2012.08.001.

Anders, Simon, and Wolfgang Huber. 2010. 'Differential Expression Analysis for Sequence Count Data.' *Genome Biology* 11 (10). BioMed Central: R106. doi:10.1186/gb-2010-11-10-r106.

Andersson, Leif, Alan L. Archibald, Cynthia D. Bottema, Rudiger Brauning, Shane C. Burgess, Dave W. Burt, Eduardo Casas, et al. 2015. 'Coordinated International Action to Accelerate Genome-to-Phenome with FAANG, the Functional Annotation of Animal Genomes Project.' *Genome Biology* 16 (1): 57. doi:10.1186/s13059-015-0622-4.

Andrée, Birgit, Tina Hillemann, Gania Kessler-Icekson, Thomas Schmitt-John, Harald Jockusch, Hans Henning Arnold, and Thomas Brand. 2000. 'Isolation and Characterization of the Novel Popeye Gene Family Expressed in Skeletal Muscle and Heart.' *Developmental Biology* 223 (2): 371–82. doi:10.1006/dbio.2000.9751.

Aran, Dvir, Zicheng Hu, and Atul J. Butte. 2017. 'XCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape.' *Genome Biology* 18 (1). BioMed Central: 220. doi:10.1186/s13059-017-1349-1.

Arendt, Maja, Tove Fall, Kerstin Lindblad-Toh, and Erik Axelsson. 2014. 'Amylase Activity Is Associated with AMY2B Copy Numbers in Dog: Implications for Dog Domestication, Diet and Diabetes.' *Animal Genetics* 45 (5). Wiley-Blackwell: 716–22. doi:10.1111/age.12179.

Bartolomei, Marisa S., Sharon Zemel, and Shirley M. Tilghman. 1991. 'Parental Imprinting of the Mouse H19 Gene.' *Nature* 351 (6322). Nature Publishing Group: 153–55. doi:10.1038/351153a0.

Batista, Pedro J, and Howard Y Chang. 2013. 'Long Noncoding RNAs: Cellular Address Codes in Development and Disease.' *Cell* 152 (6). NIH Public Access: 1298–1307. doi:10.1016/j.cell.2013.02.012.

Batzir, Nurit Assia, Mordechai Shohat, and Idit Maya. 2015. 'Chromosomal Microarray Analysis (CMA) a Clinical Diagnostic Tool in the Prenatal and Postnatal Settings.' *Pediatric Endocrinology Reviews : PER* 13 (1): 448–54.

Behjati, Sam, and Patrick S Tarpey. 2013. 'What Is next Generation Sequencing?' *Archives of Disease in Childhood. Education and Practice Edition* 98 (6). BMJ Publishing Group: 236–38. doi:10.1136/archdischild-2013-304340.

Beltran, Manuel, Isabel Puig, Cristina Peña, José Miguel García, Ana Belén Alvarez, Raúl Peña, Félix Bonilla, and Antonio García de Herreros. 2008. 'A Natural Antisense Transcript

Regulates Zeb2/Sip1 Gene Expression during Snail1-Induced Epithelial-Mesenchymal Transition.' *Genes & Development* 22 (6). Cold Spring Harbor Laboratory Press: 756–69. doi:10.1101/gad.455708.

Benson, Dennis A, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. 2013. 'GenBank.' *Nucleic Acids Research* 41 (Database issue). Oxford University Press: D36-42. doi:10.1093/nar/gks1195.

Bergman, PJ., MS Kent, and JP Farese. 2013. *Melanoma in Withrow and MacEwen's Small Animal Clinical Oncology; Page*. Saunders Elsevier.

Bertone, P., Viktor Stolc, Thomas E Royce, Joel S Rozowsky, Alexander E Urban, Xiaowei Zhu, John L Rinn, et al. 2004. 'Global Identification of Human Transcribed Sequences with Genome Tiling Arrays'. *Science* 306 (5705): 2242–46. doi:10.1126/science.1103388.

Bosenberg, Marcus, Heinz Arnheiter, and Robert Kelsh. 2014. 'Melanoma in Mankind's Best Friend.' *Pigment Cell & Melanoma Research* 27 (1): 1. doi:10.1111/pcmr.12196.

Botigué, Laura R., Shiya Song, Amelie Scheu, Shyamalika Gopalan, Amanda L. Pendleton, Matthew Oetjens, Angela M. Taravella, et al. 2017. 'Ancient European Dog Genomes Reveal Continuity since the Early Neolithic.' *Nature Communications* 8 (July): 16082. doi:10.1038/ncomms16082.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. 'Near-Optimal Probabilistic RNA-Seq Quantification.' *Nature Biotechnology* 34 (5). Nature Publishing Group: 525–27. doi:10.1038/nbt.3519.

Breiman, Leo. 2001. 'Random Forests'. *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Brent, Michael R. 2007. 'How Does Eukaryotic Gene Prediction Work?' *Nature Biotechnology* 25 (8): 883–85. doi:10.1038/nbt0807-883.

Breschi, Alessandra, Sarah Djebali, Jesse Gillis, Dmitri D. Pervouchine, Alex Dobin, Carrie A. Davis, Thomas R. Gingeras, and Roderic Guigó. 2016. 'Gene-Specific Patterns of Expression Variation across Organs and Species.' *Genome Biology* 17 (1). Genome Biology: 151. doi:10.1186/s13059-016-1008-y.

Breschi, Alessandra, Thomas R. Gingeras, and Roderic Guigó. 2017. 'Comparative Transcriptomics in Human and Mouse.' *Nature Reviews. Genetics* 18 (7). Nature Publishing Group: 425–40. doi:10.1038/nrg.2017.19.

Brockdorff, Neil, Alan Ashworth, Graham F. Kay, Penny Cooper, Sandy Smith, Veronica M. McCabe, Dominic P. Norris, Graeme D. Penny, Dipika Patel, and Sohaila Rastan. 1991. 'Conservation of Position and Exclusive Expression of Mouse Xist from the Inactive X Chromosome.' *Nature* 351 (6324). Nature Publishing Group: 329–31. doi:10.1038/351329a0.

Broeckx, Bart J G, Christophe Hitte, Frank Coopman, Geert E C Verhoeven, Sarah De Keulenaer, Ellen De Meester, Thomas Derrien, et al. 2015. 'Improved Canine Exome Designs, Featuring NcRNAs and Increased Coverage of Protein Coding Genes.' *Scientific Reports* 5 (August). Nature Publishing Group: 12810. doi:10.1038/srep12810.

Bronson, R T. 1982. 'Variation in Age at Death of Dogs of Different Sexes and Breeds.' *American Journal of Veterinary Research* 43 (11): 2057–59.

Brown, Carolyn J., Andrea Ballabio, James L. Rupert, Ronald G. Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F. Willard. 1991. 'A Gene from the Region of the Human X Inactivation Centre Is Expressed Exclusively from the Inactive X Chromosome.' *Nature* 349 (6304). Nature Publishing Group: 38–44. doi:10.1038/349038a0.

Butelli, Eugenio, Concetta Licciardello, Yang Zhang, Jianjun Liu, Steve Mackay, Paul Bailey, Giuseppe Reforgiato-Recupero, and Cathie Martin. 2012. 'Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges.' *The Plant Cell* 24 (3). American Society of Plant Biologists: 1242–55. doi:10.1105/tpc.111.095232.

Cabili, Moran N, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L. Rinn. 2011. 'Integrative Annotation of Human Large Intergenic Noncoding RNAs Reveals Global Properties and Specific Subclasses.' *Genes & Development* 25 (18). Cold Spring Harbor Laboratory Press: 1915–27. doi:10.1101/gad.17446611.

Cai, B, X Q Song, J P Cai, and S Zhang. 2014. 'HOTAIR: A Cancer-Related Long Non-Coding

RNA.' *Neoplasma* 61 (4): 379–91.

Cai, Qingyun, and Hock Peng Chan. 2016. 'A Double Application of the Benjamini-Hochberg Procedure for Testing Batched Hypotheses'. *Methodol Comput Appl Probab* 19 (March). Springer US: 429–443. doi:10.1074/mcp.M110.004374.

Cai, Xuezhong, and Bryan R Cullen. 2007. 'The Imprinted H19 Noncoding RNA Is a Primary MicroRNA Precursor.' *RNA (New York, N.Y.)* 13 (3). Cold Spring Harbor Laboratory Press: 313–16. doi:10.1261/rna.351707.

Carninci, P, T Kasukawa, S Katayama, J Gough, M C Frith, N Maeda, R Oyama, et al. 2005. 'The Transcriptional Landscape of the Mammalian Genome.' *Science (New York, N.Y.)* 309 (5740): 1559–63. doi:10.1126/science.1112014.

Casper, Jonathan, Ann S Zweig, Chris Villarreal, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, et al. 2018. 'The UCSC Genome Browser Database: 2018 Update.' *Nucleic Acids Research* 46 (D1). Oxford University Press: D762–69. doi:10.1093/nar/gkx1020.

Chandramohan, Raghu, Po-Yen Wu, John H Phan, and May D Wang. 2013. 'Benchmarking RNA-Seq Quantification Tools.' *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* 2013. NIH Public Access: 647–50. doi:10.1109/EMBC.2013.6609583.

Chang, Alfred E, Lucy Hynds Karnell, and Herman R Menck. 1998. 'The National Cancer Data Base Report on Cutaneous and Noncutaneous Melanoma: A Summary of 84,836 Cases from the Past Decade.' *Cancer* 83 (8): 1664–78. doi:10.1002/(SICI)1097-0142(19981015)83:83.0.CO;2-G.

Chen, Zhen-Hua, Wen-Tao Wang, Wei Huang, Ke Fang, Yu-Meng Sun, Shu-Rong Liu, Xue-Qun Luo, and Yue-Qin Chen. 2017. 'The LncRNA HOTAIRM1 Regulates the Degradation of PML-RARA Oncoprotein and Myeloid Cell Differentiation by Enhancing the Autophagy Pathway.' *Cell Death and Differentiation* 24 (2). Nature Publishing Group: 212–24. doi:10.1038/cdd.2016.111.

Cho, J C, and J M Tiedje. 2001. 'Bacterial Species Determination from DNA-DNA Hybridization

by Using Genome Fragments and DNA Microarrays.' *Applied and Environmental Microbiology* 67 (8). American Society for Microbiology (ASM): 3677–82. doi:10.1128/AEM.67.8.3677-3682.2001.

Chuong, Edward B, Nels C Elde, and Cédric Feschotte. 2017. 'Regulatory Activities of Transposable Elements: From Conflicts to Benefits.' *Nature Reviews. Genetics* 18 (2): 71–86. doi:10.1038/nrg.2016.139.

Clemson, C M, J A McNeil, H F Willard, and J B Lawrence. 1996. 'XIST RNA Paints the Inactive X Chromosome at Interphase: Evidence for a Novel RNA Involved in Nuclear/Chromosome Structure.' *The Journal of Cell Biology* 132 (3). The Rockefeller University Press: 259–75.

Cordonnier, A, J F Casella, and T Heidmann. 1995. 'Isolation of Novel Human Endogenous Retrovirus-like Elements with Foamy Virus-Related Pol Sequence.' *Journal of Virology* 69 (9). American Society for Microbiology (ASM): 5890–97.

Cormier, Alexandre, Komlan Avia, Lieven Sterck, Thomas Derrien, Valentin Wucher, Gwendoline Andres, Misharl Monsoor, et al. 2017. 'Re-Annotation, Improved Large-Scale Assembly and Establishment of a Catalogue of Noncoding Loci for the Genome of the Model Brown Alga Ectocarpus.' *The New Phytologist* 214 (1): 219–32. doi:10.1111/nph.14321.

Davis, Brian W., and Elaine A. Ostrander. 2014. 'Domestic Dogs and Cancer Research: A Breed-Based Genomics Approach.' *ILAR Journal* 55 (1). Oxford University Press: 59–68. doi:10.1093/ilar/ilu017.

Davis, Matthew P, Claudia Carrieri, Harpreet K Saini, Stijn van Dongen, Tommaso Leonardi, Giovanni Bussotti, Jack M Monahan, et al. 2017. 'Transposon-Driven Transcription Is a Conserved Feature of Vertebrate Spermatogenesis and Transcript Evolution.' *EMBO Reports* 18 (7): 1231–47. doi:10.15252/embr.201744059.

Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. 'Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++.' Edited by Wyeth W. Wasserman. *PLoS Computational Biology* 6 (12). Public Library of Science: e1001025.

doi:10.1371/journal.pcbi.1001025.

Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, et al. 2012. 'The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression.' *Genome Research* 22 (9): 1775–89. doi:10.1101/gr.132159.111.

Derrien, Thomas, Amaury Vaysse, Catherine André, and Christophe Hitte. 2012. 'Annotation of the Domestic Dog Genome Sequence: Finding the Missing Genes.' *Mammalian Genome : Official Journal of the International Mammalian Genome Society* 23 (1–2): 124–31. doi:10.1007/s00335-011-9372-0.

Djebali, Sarah, Valentin Wucher, Sylvain Foissac, Christophe Hitte, Evan Corre, and Thomas Derrien. 2017. 'Bioinformatics Pipeline for Transcriptome Sequencing Analysis.' *Methods in Molecular Biology (Clifton, N.J.)* 1468 (20): 201–19. doi:10.1007/978-1-4939-4035-6_14.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. 'STAR: Ultrafast Universal RNA-Seq Aligner.' *Bioinformatics (Oxford, England)* 29 (1): 15–21. doi:10.1093/bioinformatics/bts635.

Dobson, Jane M. 2013. 'Breed-Predispositions to Cancer in Pedigree Dogs.' *ISRN Veterinary Science* 2013. Hindawi Limited: 941275. doi:10.1155/2013/941275.

Egenvall, Agneta, Brenda N. Bonnett, Patrik Öhagen, Pekka Olson, Åke Hedhammar, and Henrik von Euler. 2005. 'Incidence of and Survival after Mammary Tumors in a Population of over 80,000 Insured Female Dogs in Sweden from 1995 to 2002'. *Preventive Veterinary Medicine* 69 (1–2). Elsevier: 109–27. doi:10.1016/J.PREVETMED.2005.01.014.

Emera, Deena, and Günter P Wagner. 2012. 'Transformation of a Transposon into a Derived Prolactin Promoter with Function during Human Pregnancy.' *Proceedings of the National Academy of Sciences of the United States of America* 109 (28). National Academy of Sciences: 11246–51. doi:10.1073/pnas.1118566109.

ENCODE Project Consortium. 2012. 'An Integrated Encyclopedia of DNA Elements in the

Human Genome.' *Nature* 489 (7414): 57–74. doi:10.1038/nature11247.

ENCODE Project Consortium, Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, et al. 2007. 'Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project.' *Nature* 447 (7146). NIH Public Access: 799–816. doi:10.1038/nature05874.

Feller, Liviu, Aubrey Masilana, Razia A G Khammissa, Mario Altini, Yusuf Jadwat, and Johan Lemmer. 2014. 'Melanin: The Biophysiology of Oral Melanocytes and Physiological Oral Pigmentation.' *Head & Face Medicine* 10 (March). BioMed Central: 8. doi:10.1186/1746-160X-10-8.

Franke, Vedran, Sravya Ganesh, Rosa Karlic, Radek Malik, Josef Pasulka, Filip Horvat, Maja Kuzman, et al. 2017. 'Long Terminal Repeats Power Evolution of Genes and Gene Expression Programs in Mammalian Oocytes and Zygotes.' *Genome Research* 27 (8): 1384–94. doi:10.1101/gr.216150.116.

Frantz, Laurent A.F., Victoria E Mullin, Maud Pionnier-Capitan, Ophélie Lebrasseur, Morgane Ollivier, Angela Perri, Anna Linderholm, et al. 2016. 'Genomic and Archaeological Evidence Suggests a Dual Origin of Domestic Dogs'. *Science* 352 (6290). American Association for the Advancement of Science: 1228–31. doi:10.1126/science.aaf3161.

Freeman, Kim P, Kevin A Hahn, F Dee Harris, and Glen K King. 2003. 'Treatment of Dogs with Oral Melanoma by Hypofractionated Radiation Therapy and Platinum-Based Chemotherapy (1987-1997).' *Journal of Veterinary Internal Medicine* 17 (1): 96–101.

Furney, Simon J, Samra Turajlic, Gordon Stamp, Mahrokh Nohadani, Anna Carlisle, J Meirion Thomas, Andrew Hayes, et al. 2013. 'Genome Sequencing of Mucosal Melanomas Reveals That They Are Driven by Distinct Mechanisms from Cutaneous Melanoma.' *The Journal of Pathology* 230 (3): 261–69. doi:10.1002/path.4204.

Gamalo-Siebers, Margaret, Aijun Gao, Mani Lakshminarayanan, Guanghan Liu, Fanni Natanegara, Radha Railkar, Heinz Schmidli, and Guochen Song. 2016. 'Bayesian Methods for the Design and Analysis of Noninferiority Trials.' *Journal of Biopharmaceutical Statistics* 26 (5): 823–41. doi:10.1080/10543406.2015.1074920.

Giger, Sargan, and McNiel. 2006. 'Breed-Specific Hereditary Diseases and Genetic Screening'.

Cold Spring Harbor Laboratory Press, 249–89.

Gillard, Marc, Edouard Cadieu, Clotilde De Brito, Jérôme Abadie, Béatrice Vergier, Patrick Devauchelle, Frédérique Degorce, et al. 2014. 'Naturally Occurring Melanomas in Dogs as Models for Non-UV Pathways of Human Melanomas.' *Pigment Cell & Melanoma Research* 27 (1): 90–102. doi:10.1111/pcmr.12170.

Grall, Anaïs, Eric Guaguère, Sandrine Planchais, Susanne Grond, Emmanuelle Bourrat, Ingrid Hausser, Christophe Hitte, et al. 2012. 'PNPLA1 Mutations Cause Autosomal Recessive Congenital Ichthyosis in Golden Retriever Dogs and Humans.' *Nature Genetics* 44 (2). Nature Publishing Group: 140–47. doi:10.1038/ng.1056.

Gruber, Andreas R, Stephan H Bernhart, Ivo L Hofacker, and Stefan Washietl. 2008. 'Strategies for Measuring Evolutionary Conservation of RNA Secondary Structures.' *BMC Bioinformatics* 9 (February). BioMed Central: 122. doi:10.1186/1471-2105-9-122.

GTEx Consortium. 2013. 'The Genotype-Tissue Expression (GTEx) Project.' *Nature Genetics* 45 (6). NIH Public Access: 580–85. doi:10.1038/ng.2653.

Guo, Linna, Lijie Yao, and Yang Jiang. 2016. 'A Novel Integrative Approach to Identify LncRNAs Associated with the Survival of Melanoma Patients'. *Gene* 585 (2). Elsevier B.V.: 216–20. doi:10.1016/j.gene.2016.03.036.

Guo, Yang, Ying Hu, Mingming Hu, Jiabei He, and Baolan Li. 2018. 'Long Non-Coding RNA ZEB2-AS1 Promotes Proliferation and Inhibits Apoptosis in Human Lung Cancer Cells.' *Oncology Letters* 15 (4). Spandidos Publications: 5220–26. doi:10.3892/ol.2018.7918.

Guttman, Mitchell, Ido Amit, Manuel Garber, Courtney French, Michael F Lin, David Feldser, Maite Huarte, et al. 2009. 'Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-Coding RNAs in Mammals.' *Nature* 458 (7235): 223–27. doi:10.1038/nature07672.

Gwosdz, Christian, Kathrin Scheckenbach, Oliver Lieven, Julia Reifenberger, Andreas Knopf, Henning Bier, and Vera Balz. 2006. 'Comprehensive Analysis of the P53 Status in Mucosal and Cutaneous Melanomas.' *International Journal of Cancer* 118 (3). Wiley-Blackwell: 577–82. doi:10.1002/ijc.21366.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua

Bowden, Matthew Brian Couger, et al. 2013. 'De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis.' *Nature Protocols* 8 (8). NIH Public Access: 1494–1512. doi:10.1038/nprot.2013.084.

Hagan, John P., Brittany L. O'Neill, Colin L. Stewart, Serguei V. Kozlov, and Carlo M. Croce. 2009. 'At Least Ten Genes Define the Imprinted Dlk1-Dio3 Cluster on Mouse Chromosome 12qF1.' *PloS One* 4 (2). Public Library of Science: e4352. doi:10.1371/journal.pone.0004352.

Hansen, K, and C Khanna. 2004. 'Spontaneous and Genetically Engineered Animal Models; Use in Preclinical Cancer Drug Development.' *European Journal of Cancer (Oxford, England : 1990)* 40 (6). Pergamon: 858–80. doi:10.1016/j.ejca.2003.11.031.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. 'GENCODE: The Reference Human Genome Annotation for The ENCODE Project.' *Genome Research* 22 (9): 1760–74. doi:10.1101/gr.135350.111.

Hashimshony, Tamar, Florian Wagner, Noa Sher, and Itai Yanai. 2012. 'CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification.' *Cell Reports* 2 (3). Cell Press: 666–73. doi:10.1016/j.celrep.2012.08.003.

Hayward, Jessica J., Marta G. Castelhano, Kyle C. Oliveira, Elizabeth Corey, Cheryl Balkman, Tara L. Baxter, Margret L. Casal, et al. 2016. 'Complex Disease and Phenotype Mapping in the Domestic Dog.' *Nature Communications* 7 (January). Nature Publishing Group: 10460. doi:10.1038/ncomms10460.

Hayward, Nicholas K., James S. Wilmott, Nicola Waddell, Peter A. Johansson, Matthew A. Field, Katia Nones, Ann-Marie Patch, et al. 2017. 'Whole-Genome Landscapes of Major Melanoma Subtypes.' *Nature* 545 (7653). Nature Publishing Group: 175–80. doi:10.1038/nature22071.

Hegemann, Miriam, Arnulf Stenzl, Jens Bedke, Kim N. Chi, Peter C. Black, and Tilman Todenhöfer. 2016. 'Liquid Biopsy: Ready to Guide Therapy in Advanced Prostate Cancer?' *BJU International* 118 (6): 855–63. doi:10.1111/bju.13586.

Hélias-Rodzewicz, Zofia, Elisa Funck-Brentano, Laure Baudoux, Chan Kwon Jung, Ute Zimmermann, Cristi Marin, Thierry Clerici, et al. 2015. 'Variations of BRAF Mutant Allele Percentage in Melanomas.' *BMC Cancer* 15 (1). BMC Cancer: 497. doi:10.1186/s12885-015-1515-3.

Hernandez, Belen, Hibret A Adissu, Bih-Rong Wei, Helen T Michael, Glenn Merlino, and R Mark Simpson. 2018. 'Naturally Occurring Canine Melanoma as a Predictive Comparative Oncology Model for Human Mucosal and Other Triple Wild-Type Melanomas.' *International Journal of Molecular Sciences* 19 (2). Multidisciplinary Digital Publishing Institute (MDPI). doi:10.3390/ijms19020394.

Herrero, Javier, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J Vilella, et al. 2016. 'Ensembl Comparative Genomics Resources.' *Database : The Journal of Biological Databases and Curation* 2016 (November). Oxford University Press: bav096. doi:10.1093/database/bav096.

Hirsch, Cory D., and Nathan M. Springer. 2017. 'Transposable Element Influences on Gene Expression in Plants.' *Biochimica et Biophysica Acta* 1860 (1). Elsevier: 157–65. doi:10.1016/j.bbagrm.2016.05.010.

Hitte, Christophe, Jennifer Madeoy, Ewen F. Kirkness, Catherine Priat, Travis D. Lorentzen, Fabrice Senger, Dan Thomas, et al. 2005. 'Facilitating Genome Navigation: Survey Sequencing and Dense Radiation-Hybrid Gene Mapping.' *Nature Reviews. Genetics* 6 (8): 643–48. doi:10.1038/nrg1658.

Hoeppner, Marc P., Andrew Lundquist, Mono Pirun, Jennifer R S Meadows, Neda Zamani, Jeremy Johnson, Görel Sundström, et al. 2014. 'An Improved Canine Genome and a Comprehensive Catalogue of Coding Genes and Non-Coding Transcripts.' *PloS One* 9 (3): e91172. doi:10.1371/journal.pone.0091172.

Hsieh, Ricardo, Marcello M. S. Nico, Claudia M Coutinho-Camillo, Marcilei E. Buim, Martin Sangueza, and Silvia V Lourenço. 2013. 'The CDKN2A and MAP Kinase Pathways: Molecular Roads to Primary Oral Mucosal Melanoma.' *The American Journal of Dermatopathology* 35 (2). The American Journal of Dermatopathology: 167–75. doi:10.1097/DAD.0b013e31825fa1f6.

Ilie, Marius, Élodie Long, Véronique Hofman, Virginie Lespinet, Olivier Bordone, Kevin Washetine, Virginie Gavric-Tanga, and Paul Hofman. 2014. 'Les Méthodes de Séquençage de « nouvelle Génération » (NGS) et Le Cancer Broncho-Pulmonaire: Principales Technologies, Applications et Limites Actuelles En Pathologie'. *Revue Francophone Des Laboratoires* 2014 (458). Elsevier Masson: 51–58. doi:10.1016/S1773-035X(14)72315-9.

Institut national du Cancer. 2017. 'Les Cancers En France'.

Johnson, Rory, and Roderic Guigó. 2014. 'The RIDL Hypothesis: Transposable Elements as Functional Domains of Long Noncoding RNAs.' *RNA (New York, N.Y.)* 20 (7): 959–76. doi:10.1261/rna.044560.114.

Kanduri, Chandrasekhar. 2016. 'Long Noncoding RNAs: Lessons from Genomic Imprinting.' *Biochimica et Biophysica Acta* 1859 (1). Elsevier: 102–11. doi:10.1016/j.bbagrm.2015.05.006.

Kapitonov, Vladimir V, and Eugene V Koonin. 2015. 'Evolution of the RAG1-RAG2 Locus: Both Proteins Came from the Same Transposon.' *Biology Direct* 10 (April). BioMed Central: 20. doi:10.1186/s13062-015-0055-8.

Kapranov, P., J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, et al. 2007. 'RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription'. *Science* 316 (5830): 1484–88. doi:10.1126/science.1138341.

Kapusta, Aurélie, Zev Kronenberg, Vincent J. Lynch, Xiaoyu Zhuo, LeeAnn Ramsay, Guillaume Bourque, Mark Yandell, and Cédric Feschotte. 2013. 'Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs.' *PLoS Genetics* 9 (4): e1003470. doi:10.1371/journal.pgen.1003470.

Ke, Dong, Hanwei Li, Yi Zhang, Yinghong An, Hanjiang Fu, Xuedong Fang, and Xiaofei Zheng. 2017. 'The Combination of Circulating Long Noncoding RNAs AK001058, INHBA-AS1, MIR4435-2HG, and CEBPA-AS1 Fragments in Plasma Serve as Diagnostic Markers for Gastric Cancer.' *Oncotarget* 8 (13). Impact Journals, LLC: 21516–25. doi:10.18632/oncotarget.15628.

Ke, X., L. J. Kennedy, A. D. Short, E. H. Seppälä, A. Barnes, D. N. Clements, S. H. Wood, et al.

2011. ‘Assessment of the Functionality of Genome-Wide Canine SNP Arrays and Implications for Canine Disease Association Studies.’ *Animal Genetics* 42 (2). Wiley/Blackwell (10.1111): 181–90. doi:10.1111/j.1365-2052.2010.02132.x.

Kelley, David, and John Rinn. 2012. ‘Transposable Elements Reveal a Stem Cell-Specific Class of Long Noncoding RNAs.’ *Genome Biology* 13 (11): R107. doi:10.1186/gb-2012-13-11-r107.

Kerboua, E., C. Rancoule, N. Magné, and K. Bouzid. 2016. ‘Récepteur de l’epidermal Growth Factor (EGFR), Marqueur Pronostique Des Carcinomes Indifférenciés Du Nasopharynx : Étude Nord-Africaine’. *Cancer/Radiothérapie* 20 (6–7). Elsevier Masson: 716–17. doi:10.1016/J.CANRAD.2016.08.011.

Khaitan, Divya, Marcel E Dinger, Joseph Mazar, Joanna Crawford, Martin A Smith, John S Mattick, and Ranjan J Perera. 2011. ‘The Melanoma-Upregulated Long Noncoding RNA SPRY4-IT1 Modulates Apoptosis and Invasion.’ *Cancer Research* 71 (11). American Association for Cancer Research: 3852–62. doi:10.1158/0008-5472.CAN-10-4460.

Khaitovich, Philipp, Janet Kelso, Henriette Franz, Johann Visagie, Thomas Giger, Sabrina Joerchel, Ekkehard Petzold, Richard E Green, Michael Lachmann, and Svante Pääbo. 2006. ‘Functionality of Intergenic Transcription: An Evolutionary Comparison.’ *PLoS Genetics* 2 (10). Public Library of Science: e171. doi:10.1371/journal.pgen.0020171.

Kharchenko, Peter V, Lev Silberstein, and David T Scadden. 2014. ‘Bayesian Approach to Single-Cell Differential Expression Analysis.’ *Nature Methods* 11 (7). Nature Publishing Group: 740–42. doi:10.1038/nmeth.2967.

Kirkness, Ewen F, Vineet Bafna, Aaron L Halpern, Samuel Levy, Karin Remington, Douglas B Rusch, Arthur L Delcher, et al. 2003. ‘The Dog Genome: Survey Sequencing and Comparative Analysis.’ *Science (New York, N.Y.)* 301 (5641). American Association for the Advancement of Science: 1898–1903. doi:10.1126/science.1086432.

Kladi-Skandali, Athina, Kleita Michaelidou, Andreas Scorilas, and Konstantinos Mavridis. 2015. ‘Long Noncoding RNAs in Digestive System Malignancies: A Novel Class of Cancer Biomarkers and Therapeutic Targets?’ *Gastroenterology Research and Practice* 2015. Hindawi Limited: 319861. doi:10.1155/2015/319861.

Kornienko, Aleksandra E., Christoph P. Dotter, Philipp M. Guenzl, Heinz Gisslinger, Bettina Gisslinger, Ciara Cleary, Robert Kralovics, Florian M. Pauler, and Denise P. Barlow. 2016. 'Long Non-Coding RNAs Display Higher Natural Expression Variation than Protein-Coding Genes in Healthy Humans.' *Genome Biology* 17 (1). Genome Biology: 14. doi:10.1186/s13059-016-0873-8.

Kretschmer, Alexander, and Derya Tilki. 2017. 'Biomarkers in Prostate Cancer - Current Clinical Utility and Future Perspectives.' *Critical Reviews in Oncology/Hematology* 120 (December): 180–93. doi:10.1016/j.critrevonc.2017.11.007.

Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. 2009. 'Circos: An Information Aesthetic for Comparative Genomics.' *Genome Research* 19 (9). Cold Spring Harbor Laboratory Press: 1639–45. doi:10.1101/gr.092759.109.

Kutter, Claudia, Stephen Watt, Klara Stefflova, Michael D. Wilson, Angela Goncalves, Chris P. Ponting, Duncan T. Odom, and Ana C. Marques. 2012. 'Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression.' *PLoS Genetics* 8 (7): e1002841. doi:10.1371/journal.pgen.1002841.

Lagarde, Julien, Barbara Uszczynska-Ratajczak, Silvia Carbonell, Sílvia Pérez-Lluch, Amaya Abad, Carrie Davis, Thomas R. Gingeras, et al. 2017. 'High-Throughput Annotation of Full-Length Long Noncoding RNAs with Capture Long-Read Sequencing.' *Nature Genetics* 49 (12). Nature Publishing Group: 1731–40. doi:10.1038/ng.3988.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. 'Initial Sequencing and Analysis of the Human Genome.' *Nature* 409 (6822): 860–921. doi:10.1038/35057062.

Langfelder, Peter, and Steve Horvath. 2008. 'WGCNA: An R Package for Weighted Correlation Network Analysis'. *BMC Bioinformatics* 9 (1). BioMed Central: 559. doi:10.1186/1471-2105-9-559.

Le Béguec, Céline, Valentin Wucher, Lætitia Lagoutte, Edouard Cadieu, Nadine Botherel, Benoît Hédan, Clotilde De Brito, et al. 2018. 'Characterisation and Functional Predictions of Canine Long Non-Coding RNAs.' *Scientific Reports* 8 (1). Nature Publishing

Group: 13444. doi:10.1038/s41598-018-31770-2.

Leahy, Michael F, John F Seymour, Rodney J Hicks, and J Harvey Turner. 2006. 'Multicenter Phase II Clinical Study of Iodine-131-Rituximab Radioimmunotherapy in Relapsed or Refractory Indolent Non-Hodgkin's Lymphoma.' *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 24 (27): 4418–25. doi:10.1200/JCO.2005.05.3470.

Leng, Ning, John A. Dawson, James A. Thomson, Victor Ruotti, Anna I. Rissman, Bart M G Smits, Jill D. Haag, Michael N. Gould, Ron M. Stewart, and Christina Kendziorski. 2013. 'EBSeq: An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments.' *Bioinformatics (Oxford, England)* 29 (8): 1035–43. doi:10.1093/bioinformatics/btt087.

Lequarré, Anne-Sophie, Leif Andersson, Catherine André, Merete Fredholm, Christophe Hitte, Tosso Leeb, Hannes Lohi, Kerstin Lindblad-Toh, and Michel Georges. 2011. 'LUPA: A European Initiative Taking Advantage of the Canine Genome Architecture for Unravelling Complex Disorders in Both Human and Dogs.' *Veterinary Journal (London, England : 1997)* 189 (2). W.B. Saunders: 155–59. doi:10.1016/j.tvjl.2011.06.013.

Lessard, Laurent, Michelle Liu, Diego M Marzese, Hongwei Wang, Kelly Chong, Neal Kawas, Nicholas C Donovan, et al. 2015. 'The CASC15 Long Intergenic Noncoding RNA Locus Is Involved in Melanoma Progression and Phenotype Switching.' *The Journal of Investigative Dermatology* 135 (10). NIH Public Access: 2464–74. doi:10.1038/jid.2015.200.

Leucci, Eleonora, Elizabeth A. Coe, Jean-Christophe Marine, and Keith W. Vance. 2016. 'The Emerging Role of Long Non-Coding RNAs in Cutaneous Melanoma.' *Pigment Cell & Melanoma Research* 29 (6): 619–26. doi:10.1111/pcmr.12537.

Leucci, Eleonora, Roberto Vendramin, Marco Spinazzi, Patrick Laurette, Mark Fiers, Jasper Wouters, Enrico Radaelli, et al. 2016. 'Melanoma Addiction to the Long Non-Coding RNA SAMMSON.' *Nature* 531 (7595). Nature Publishing Group: 518–22. doi:10.1038/nature17161.

Li, Bo, and Colin N. Dewey. 2011. 'RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.' *BMC Bioinformatics* 12 (August): 323.

doi:10.1186/1471-2105-12-323.

Li, Jun, and Robert Tibshirani. 2013. 'Finding Consistent Patterns: A Nonparametric Approach for Identifying Differential Expression in RNA-Seq Data.' *Statistical Methods in Medical Research* 22 (5). NIH Public Access: 519–36. doi:10.1177/0962280211428386.

Li, Ling, Tingting Feng, Yingying Lian, Guangfeng Zhang, Alan Garen, and Xu Song. 2009. 'Role of Human Noncoding RNAs in the Control of Tumorigenesis.' *Proceedings of the National Academy of Sciences of the United States of America* 106 (31). National Academy of Sciences: 12956–61. doi:10.1073/pnas.0906005106.

Li, Ruiya, Lingli Zhang, Lizhou Jia, Yan Duan, Yan Li, Lidao Bao, and Na Sha. 2014. 'Long Non-Coding RNA BANCR Promotes Proliferation in Malignant Melanoma by Regulating MAPK Pathway Activation.' *PloS One* 9 (6). Public Library of Science: e100893. doi:10.1371/journal.pone.0100893.

Li, T, J Xie, C Shen, D Cheng, Y Shi, Z Wu, X Deng, et al. 2016. 'Upregulation of Long Noncoding RNA ZEB1-AS1 Promotes Tumor Metastasis and Predicts Poor Prognosis in Hepatocellular Carcinoma.' *Oncogene* 35 (12). Nature Publishing Group: 1575–84. doi:10.1038/onc.2015.223.

Li, Xiaolei, Zhiqiang Wu, Xiaobing Fu, and Weidong Han. 2014. 'LncRNAs: Insights into Their Function and Mechanics in Underlying Disorders.' *Mutation Research. Reviews in Mutation Research* 762 (October). Elsevier: 1–21. doi:10.1016/j.mrrev.2014.04.002.

Li, Xing, Mengning Zheng, Jinding Pu, Yumin Zhou, Hui Pan, Wei Hong, Xin Fu, et al. 2018. 'Identification of Abnormally Expressed LncRNAs Induced by PM2.5 in Human Bronchial Epithelial Cells.' *Bioscience Reports*, June, BSR20171577. doi:10.1042/BSR20171577.

Lindblad-Toh, Kerstin, Claire M. Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, et al. 2005. 'Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog.' *Nature* 438 (7069): 803–19. doi:10.1038/nature04338.

Lisch, Damon. 2013. 'How Important Are Transposons for Plant Evolution?' *Nature Reviews. Genetics* 14 (1): 49–61. doi:10.1038/nrg3374.

Liu, Y-w, M Sun, R Xia, E-b Zhang, X-h Liu, Z-h Zhang, T-p Xu, W De, B-r Liu, and Z-x Wang.

2015. 'LincHOTAIR Epigenetically Silences MiR34a by Binding to PRC2 to Promote the Epithelial-to-Mesenchymal Transition in Human Gastric Cancer.' *Cell Death & Disease* 6 (7). Nature Publishing Group: e1802. doi:10.1038/cddis.2015.150.

Lizio, Marina, Abdul Kadir Mukarram, Mizuho Ohno, Shoko Watanabe, Masayoshi Itoh, Akira Hasegawa, Timo Lassmann, et al. 2017. 'Monitoring Transcription Initiation Activities in Rat and Dog.' *Scientific Data* 4 (November). Nature Publishing Group: 170173. doi:10.1038/sdata.2017.173.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. 'Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.' *Genome Biology* 15 (12). BioMed Central: 550. doi:10.1186/s13059-014-0550-8.

Lu, Shan, Jiansheng Zhou, Yimin Sun, Nan Li, Mingyong Miao, Binghua Jiao, and Huan Chen. 2017. 'The Noncoding RNA HOXD-AS1 Is a Critical Regulator of the Metastasis and Apoptosis Phenotype in Human Hepatocellular Carcinoma.' *Molecular Cancer* 16 (1). BioMed Central: 125. doi:10.1186/s12943-017-0676-x.

Lubelsky, Yoav, and Igor Ulitsky. 2018. 'Sequences Enriched in Alu Repeats Drive Nuclear Localization of Long RNAs in Human Cells.' *Nature* 555 (7694). Nature Publishing Group: 107–11. doi:10.1038/nature25757.

Luk, Alfred Chun-Shui, Wai-Yee Chan, Owen M Rennert, and Tin-Lap Lee. 2014. 'Long Noncoding RNAs in Spermatogenesis: Insights from Recent High-Throughput Transcriptome Studies.' *Reproduction (Cambridge, England)* 147 (5): R131-41. doi:10.1530/REP-13-0594.

MacDonald, Valerie S, Douglas H Thamm, Ilene D Kurzman, Michelle M Turek, and David M Vail. 2005. 'Does L-Asparaginase Influence Efficacy or Toxicity When Added to a Standard CHOP Protocol for Dogs with Lymphoma?' *Journal of Veterinary Internal Medicine* 19 (5): 732–36.

MacEwen, E G, A K Patnaik, H J Harvey, A A Hayes, and R Matus. 1986. 'Canine Oral Melanoma: Comparison of Surgery versus Surgery plus Corynebacterium Parvum.' *Cancer Investigation* 4 (5): 397–402.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa

Goldman, Itay Tirosh, et al. 2015. 'Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.' *Cell* 161 (5). Cell Press: 1202–14. doi:10.1016/j.cell.2015.05.002.

Mallory, Allison C, and Alena Shkumatava. 2015. 'LncRNAs in Vertebrates: Advances and Challenges.' *Biochimie* 117 (October): 3–14. doi:10.1016/j.biochi.2015.03.014.

Marcus, David M, Rebecca P Marcus, Roshan S Prabhu, Taofeek K Owonikoko, David H Lawson, Jeffrey Switchenko, and Jonathan J Beitler. 2012. 'Rising Incidence of Mucosal Melanoma of the Head and Neck in the United States.' *Journal of Skin Cancer* 2012 (December). Hindawi: 231693. doi:10.1155/2012/231693.

Mattick, John S. 2009. 'The Genetic Signatures of Noncoding RNAs.' *PLoS Genetics* 5 (4). Public Library of Science: e1000459. doi:10.1371/journal.pgen.1000459.

McClintock, B. 1956. 'Intranuclear Systems Controlling Gene Action and Mutation.' *Brookhaven Symposia in Biology*, no. 8 (February): 58–74.

Mehravar, Majid, Meisam Jafarzadeh, Maryam Kay, Hadi Najafi, Fahimeh Hosseini, Seyed Javad Mowla, and Bahram M. Soltani. 2017. 'Introduction of Novel Splice Variants for CASC18 Gene and Its Relation to the Neural Differentiation.' *Gene* 603 (March). Elsevier: 27–33. doi:10.1016/j.gene.2016.12.008.

Melé, Marta, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, et al. 2015. 'Human Genomics. The Human Transcriptome across Tissues and Individuals.' *Science (New York, N.Y.)* 348 (6235): 660–65. doi:10.1126/science.aaa0355.

Melo, Carlos A., Jarno Drost, Patrick J. Wijchers, Harmen van de Werken, Elzo de Wit, Joachim A F Oude Vrielink, Ran Elkon, et al. 2013. 'ERNAs Are Required for P53-Dependent Enhancer Activity and Gene Transcription.' *Molecular Cell* 49 (3): 524–35. doi:10.1016/j.molcel.2012.11.021.

Mercer, Tim R., Marcel E. Dinger, and John S. Mattick. 2009. 'Long Non-Coding RNAs: Insights into Functions.' *Nature Reviews. Genetics* 10 (3): 155–59. doi:10.1038/nrg2521.

Mezlini, Aziz M, Eric J M Smith, Marc Fiume, Orion Buske, Gleb L Savich, Sohrab Shah, Sam Aparicio, Derek Y Chiang, Anna Goldenberg, and Michael Brudno. 2013. 'IReckon:

Simultaneous Isoform Discovery and Abundance Estimation from RNA-Seq Data.' *Genome Research* 23 (3). Cold Spring Harbor Laboratory Press: 519–29. doi:10.1101/gr.142232.112.

Mikkelsen, Lauge H, Ann-Cathrine Larsen, Christian von Buchwald, Krzysztof T Drzewiecki, Jan U Prause, and Steffen Heegaard. 2016. 'Mucosal Malignant Melanoma - a Clinical, Oncological, Pathological and Genetic Survey.' *APMIS : Acta Pathologica, Microbiologica, et Immunologica Scandinavica* 124 (6): 475–86. doi:10.1111/apm.12529.

Mochizuki, Hiroyuki, Katherine Kennedy, Susan G Shapiro, and Matthew Breen. 2015. 'BRAF Mutations in Canine Cancers.' *PloS One* 10 (6). Public Library of Science: e0129534. doi:10.1371/journal.pone.0129534.

Montgomery, Stephen B, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. 2010. 'Transcriptome Genetics Using Second Generation Sequencing in a Caucasian Population.' *Nature* 464 (7289). Europe PMC Funders: 773–77. doi:10.1038/nature08903.

Murakami, A, T Mori, H Sakai, M Murakami, T Yanai, Y Hoshino, and K Maruo. 2011. 'Analysis of KIT Expression and KIT Exon 11 Mutations in Canine Oral Malignant Melanomas.' *Veterinary and Comparative Oncology* 9 (3): 219–24. doi:10.1111/j.1476-5829.2010.00253.x.

Muret, Kévin, Christophe Klopp, Valentin Wucher, Diane Esquerré, Fabrice Legeai, Frédéric Lecerf, Colette Désert, et al. 2017. 'Long Noncoding RNA Repertoire in Chicken Liver and Adipose Tissue.' *Genetics, Selection, Evolution : GSE* 49 (1). BioMed Central: 6. doi:10.1186/s12711-016-0275-0.

Nakagawa, Shinichi, Takao Naganuma, Go Shioi, and Tetsuro Hirose. 2011. 'Paraspeckles Are Subpopulation-Specific Nuclear Bodies That Are Not Essential in Mice.' *The Journal of Cell Biology* 193 (1). Rockefeller University Press: 31–39. doi:10.1083/jcb.201011110.

Nicolae, Marius, Serghei Mangul, Ion I Măndoiu, and Alex Zelikovsky. 2011. 'Estimation of Alternative Splicing Isoform Frequencies from RNA-Seq Data.' *Algorithms for Molecular*

Biology : AMB 6 (1). BioMed Central: 9. doi:10.1186/1748-7188-6-9.

Ollivier, Morgane, Anne Tresset, Fabiola Bastian, Laetitia Lagoutte, Erik Axelsson, Maja-Louise Arendt, Adrian Bălăşescu, et al. 2016. 'Amy2B Copy Number Variation Reveals Starch Diet Adaptations in Ancient European Dogs.' *Royal Society Open Science* 3 (11). The Royal Society: 160449. doi:10.1098/rsos.160449.

Paoloni, Melissa, and Chand Khanna. 2008. 'Translation of New Cancer Treatments from Pet Dogs to Humans.' *Nature Reviews. Cancer* 8 (2): 147–56. doi:10.1038/nrc2273.

Parker, Heidi G, Lisa V Kim, Nathan B Sutter, Scott Carlson, Travis D Lorentzen, Tiffany B Malek, Gary S Johnson, Hawkins B DeFrance, Elaine A Ostrander, and Leonid Kruglyak. 2004. 'Genetic Structure of the Purebred Domestic Dog.' *Science (New York, N.Y.)* 304 (5674): 1160–64. doi:10.1126/science.1097406.

Paterson, Clare, Yanhong Wang, Thomas M. Hyde, Daniel R. Weinberger, Joel E. Kleinman, and Amanda J. Law. 2017. 'Temporal, Diagnostic, and Tissue-Specific Regulation of NRG3 Isoform Expression in Human Brain Development and Affective Disorders.' *The American Journal of Psychiatry* 174 (3): 256–65. doi:10.1176/appi.ajp.2016.16060721.

Patrick, Ronald J., Neil A. Fenske, and Jane L. Messina. 2007. 'Primary Mucosal Melanoma.' *Journal of the American Academy of Dermatology* 56 (5). Mosby: 828–34. doi:10.1016/j.jaad.2006.06.017.

Patro, Rob, Geet Duggal, and Carl Kingsford. 2015. 'Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-Seq Data Using Lightweight-Alignment'. *BioRxiv*, June. Cold Spring Harbor Laboratory, 021592. doi:10.1101/021592.

Patro, Rob, Stephen M Mount, and Carl Kingsford. 2014. 'Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms.' *Nature Biotechnology* 32 (5). NIH Public Access: 462–64. doi:10.1038/nbt.2862.

Pendleton, Amanda L, Feichen Shen, Angela M Taravella, Sarah Emery, Krishna R Veeramah, Adam R Boyko, and Jeffrey M Kidd. 2018. 'Comparison of Village Dog and Wolf Genomes Highlights the Role of the Neural Crest in Dog Domestication.' *BMC Biology* 16 (1). BioMed Central: 64. doi:10.1186/s12915-018-0535-2.

Penso-Dolfin, Luca, Ross Swofford, Jeremy Johnson, Jessica Alföldi, Kerstin Lindblad-Toh,

David Swarbreck, Simon Moxon, and Federica Di Palma. 2016. 'An Improved MicroRNA Annotation of the Canine Genome.' *PloS One* 11 (4). Public Library of Science: e0153453. doi:10.1371/journal.pone.0153453.

Pervouchine, Dmitri D., Sarah Djebali, Alessandra Breschi, Carrie A. Davis, Pablo Prieto Barja, Alex Dobin, Andrea Tanzer, et al. 2015. 'Enhanced Transcriptome Maps from Multiple Mouse Tissues Reveal Evolutionary Constraint in Gene Expression.' *Nature Communications* 6 (January): 5903. doi:10.1038/ncomms6903.

Picelli, Simone, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2013. 'Smart-Seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells.' *Nature Methods* 10 (11). Nature Publishing Group: 1096–98. doi:10.1038/nmeth.2639.

Plassais, Jocelyn, Laetitia Lagoutte, Solenne Correard, Manon Paradis, Eric Guaguère, Benoit Hédan, Alix Pommier, et al. 2016. 'A Point Mutation in a LincRNA Upstream of GDNF Is Associated to a Canine Insensitivity to Pain: A Spontaneous Model for Human Sensory Neuropathies.' *PLoS Genetics* 12 (12): e1006482. doi:10.1371/journal.pgen.1006482.

Ponting, Chris P., Peter L. Oliver, and Wolf Reik. 2009. 'Evolution and Functions of Long Noncoding RNAs.' *Cell* 136 (4). Elsevier Inc.: 629–41. doi:10.1016/j.cell.2009.02.006.

Poorman, Kelsey, Luke Borst, Scott Moroff, Siddharth Roy, Philippe Labelle, Alison Motsinger-Reif, and Matthew Breen. 2015. 'Comparative Cytogenetic Characterization of Primary Canine Melanocytic Lesions Using Array CGH and Fluorescence in Situ Hybridization.' *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 23 (2). NIH Public Access: 171–86. doi:10.1007/s10577-014-9444-6.

Proulx, David R, David M Ruslander, Richard K Dodge, Marlene L Hauck, Laurel E Williams, Birgitte Horn, G Sylvester Price, and Donald E Thrall. 2003. 'A Retrospective Analysis of 140 Dogs with Oral Melanoma Treated with External Beam Radiation.' *Veterinary Radiology & Ultrasound : The Official Journal of the American College of Veterinary Radiology and the International Veterinary Radiology Association* 44 (3): 352–59.

Quinlan, Aaron R., and Ira M. Hall. 2010. 'BEDTools: A Flexible Suite of Utilities for Comparing

Genomic Features'. *Bioinformatics* 26 (6): 841–42. doi:10.1093/bioinformatics/btq033.

Raut, Satish K., and Madhu Khullar. 2018. 'The Big Entity of New RNA World: Long Non-Coding RNAs in Microvascular Complications of Diabetes.' *Frontiers in Endocrinology* 9 (June). Frontiers: 300. doi:10.3389/fendo.2018.00300.

Renwick, J H. 1971. 'The Mapping of Human Chromosomes.' *Annual Review of Genetics* 5 (1). Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA: 81–120. doi:10.1146/annurev.ge.05.120171.000501.

Richardson, Sandra R, Aurélien J Doucet, Huira C Kopera, John B Moldovan, José Luis Garcia-Perez, and John V Moran. 2015. 'The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes.' *Microbiology Spectrum* 3 (2). NIH Public Access: MDNA3-0061-2014. doi:10.1128/microbiolspec.MDNA3-0061-2014.

Roberts, Adam, and Lior Pachter. 2013. 'Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments.' *Nature Methods* 10 (1). NIH Public Access: 71–73. doi:10.1038/nmeth.2251.

Roberts, Adam, Harold Pimentel, Cole Trapnell, and Lior Pachter. 2011. 'Identification of Novel Transcripts in Annotated Genomes Using RNA-Seq.' *Bioinformatics (Oxford, England)* 27 (17): 2325–29. doi:10.1093/bioinformatics/btr355.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. 'EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.' *Bioinformatics (Oxford, England)* 26 (1). Oxford University Press: 139–40. doi:10.1093/bioinformatics/btp616.

Rohart, Florian, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. 2017. 'MixOmics: An R Package for 'omics Feature Selection and Multiple Data Integration.' Edited by Dina Schneidman. *PLoS Computational Biology* 13 (11): e1005752. doi:10.1371/journal.pcbi.1005752.

Rowell, Jennie L, Donna O McCarthy, and Carlos E Alvarez. 2011. 'Dog Models of Naturally Occurring Cancer.' *Trends in Molecular Medicine* 17 (7). NIH Public Access: 380–88. doi:10.1016/j.molmed.2011.02.004.

Ru, G, B Terracini, and L T Glickman. 1998. 'Host Related Risk Factors for Canine

Osteosarcoma.' *Veterinary Journal (London, England : 1997)* 156 (1): 31–39.

Saba, Corey, Melissa Paoloni, Christina Mazcko, William Kisseberth, Jenna H Burton, Annette Smith, Heather Wilson-Robles, et al. 2016. 'A Comparative Oncology Study of Iniparib Defines Its Pharmacokinetic Profile and Biological Activity in a Naturally-Occurring Canine Cancer Model.' *PloS One* 11 (2). Public Library of Science: e0149194. doi:10.1371/journal.pone.0149194.

Salameh, Ahmad, Alessandro K Lee, Marina Cardó-Vila, Diana N Nunes, Eleni Efstathiou, Fernanda I Staquicini, Andrey S Dobroff, et al. 2015. 'PRUNE2 Is a Human Prostate Cancer Suppressor Regulated by the Intronic Long Noncoding RNA PCA3.' *Proceedings of the National Academy of Sciences of the United States of America* 112 (27). National Academy of Sciences: 8403–8. doi:10.1073/pnas.1507882112.

Sanger, F, S Nicklen, and A R Coulson. 1977. 'DNA Sequencing with Chain-Terminating Inhibitors.' *Proceedings of the National Academy of Sciences of the United States of America* 74 (12). National Academy of Sciences: 5463–67.

Schena, M, D Shalon, R W Davis, and P O Brown. 1995. 'Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.' *Science (New York, N.Y.)* 270 (5235): 467–70.

Schiffman, Joshua D, and Matthew Breen. 2015. 'Comparative Oncology: What Dogs and Other Species Can Teach Us about Humans with Cancer.' *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1673). The Royal Society. doi:10.1098/rstb.2014.0231.

Schmidt, Dominic, Petra C Schwalie, Michael D Wilson, Benoit Ballester, Angela Gonçalves, Claudia Kutter, Gordon D Brown, Aileen Marshall, Paul Flicek, and Duncan T Odom. 2012. 'Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages.' *Cell* 148 (1–2). Elsevier: 335–48. doi:10.1016/j.cell.2011.11.058.

Sedgwick, Philip. 2014. 'Multiple Hypothesis Testing and Bonferroni's Correction.' *BMJ (Clinical Research Ed.)* 349 (October). British Medical Journal Publishing Group: g6284. doi:10.1136/BMJ.G6284.

Seiler, Jana, Marco Breinig, Maïwen Caudron-Herger, Maria Polycarpou-Schwarz, Michael Boutros, and Sven Diederichs. 2017. 'The LncRNA VELUCT Strongly Regulates Viability of Lung Cancer Cells despite Its Extremely Low Abundance.' *Nucleic Acids Research* 45 (9): 5458–69. doi:10.1093/nar/gkx076.

Serviss, Jason T, Per Johnsson, and Dan Grandér. 2014. 'An Emerging Role for Long Non-Coding RNAs in Cancer Metastasis.' *Frontiers in Genetics* 5. Frontiers Media SA: 234. doi:10.3389/fgene.2014.00234.

Seyednasrollah, Fatemeh, Asta Laiho, and Laura L Elo. 2015. 'Comparison of Software Packages for Detecting Differential Expression in RNA-Seq Studies.' *Briefings in Bioinformatics* 16 (1). Oxford University Press: 59–70. doi:10.1093/bib/bbt086.

Shi, Xuefei, Ming Sun, Hongbing Liu, Yanwen Yao, and Yong Song. 2013. 'Long Non-Coding RNAs: A New Frontier in the Study of Human Diseases.' *Cancer Letters* 339 (2). Elsevier: 159–66. doi:10.1016/j.canlet.2013.06.013.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. 2017. 'Cancer Statistics, 2017.' *CA: A Cancer Journal for Clinicians* 67 (1): 7–30. doi:10.3322/caac.21387.

Signal, Bethany, Brian S. Gloss, and Marcel E. Dinger. 2016. 'Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs.' *Trends in Genetics : TIG* 32 (10). Elsevier Ltd: 620–37. doi:10.1016/j.tig.2016.08.004.

Smedley, R. C., W. L. Spangler, D. G. Esplin, B. E. Kitchell, P. J. Bergman, H.-Y. Ho, I. L. Bergin, and M. Kiupel. 2011. 'Prognostic Markers for Canine Melanocytic Neoplasms: A Comparative Review of the Literature and Goals for Future Investigation.' *Veterinary Pathology* 48 (1). SAGE PublicationsSage CA: Los Angeles, CA: 54–72. doi:10.1177/0300985810390717.

Smit, AFA, Hubley, R & Green, P. 2013. 'RepeatMasker Open-4.0.'

Smyth, Gordon K. 2004. 'Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.' *Statistical Applications in Genetics and Molecular Biology* 3 (1): Article3. doi:10.2202/1544-6115.1027.

Soumillon, Magali, Anamaria Necsulea, Manuela Weier, David Brawand, Xiaolan Zhang, Hongcang Gu, Pauline Barthès, et al. 2013. 'Cellular Source and Mechanisms of High

Transcriptome Complexity in the Mammalian Testis.' *Cell Reports* 3 (6). Cell Press: 2179–90. doi:10.1016/j.celrep.2013.05.031.

Starkey, Mike P., Timothy J. Scase, Cathryn S. Mellersh, and Sue Murphy. 2005. 'Dogs Really Are Man's Best Friend--Canine Genomics Has Applications in Veterinary and Human Medicine!' *Briefings in Functional Genomics & Proteomics* 4 (2): 112–28. doi:10.1093/bfgp/4.2.112.

Sutter, Nathan B., and Elaine A. Ostrander. 2004. 'Dog Star Rising: The Canine Genetic System.' *Nature Reviews. Genetics* 5 (12). Nature Publishing Group: 900–910. doi:10.1038/nrg1492.

Sutter, Nathan B, Michael A Eberle, Heidi G Parker, Barbara J Pullar, Ewen F Kirkness, Leonid Kruglyak, and Elaine A Ostrander. 2004. 'Extensive and Breed-Specific Linkage Disequilibrium in Canis Familiaris.' *Genome Research* 14 (12). Cold Spring Harbor Laboratory Press: 2388–96. doi:10.1101/gr.3147604.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, et al. 2009. 'MRNA-Seq Whole-Transcriptome Analysis of a Single Cell.' *Nature Methods* 6 (5). Nature Publishing Group: 377–82. doi:10.1038/nmeth.1315.

Tarazona, Sonia, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. 2011. 'Differential Expression in RNA-Seq: A Matter of Depth.' *Genome Research* 21 (12). Cold Spring Harbor Laboratory Press: 2213–23. doi:10.1101/gr.124321.111.

Teng, Mingxiang, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, et al. 2016. 'A Benchmark for RNA-Seq Quantification Pipelines.' *Genome Biology* 17 (1). Genome Biology: 74. doi:10.1186/s13059-016-0940-1.

Théon, A P, C Rodriguez, and B R Madewell. 1997. 'Analysis of Prognostic Factors and Patterns of Failure in Dogs with Malignant Oral Tumors Treated with Megavoltage Irradiation.' *Journal of the American Veterinary Medical Association* 210 (6): 778–84.

Thomas, Rachael, Shannon E Duke, Huixia J Wang, Tessa E Breen, Robert J Higgins, Keith E Linder, Peter Ellis, et al. 2009. '"Putting Our Heads Together": Insights into Genomic Conservation between Human and Canine Intracranial Tumors.' *Journal of Neuro-Oncology* 94 (3). NIH Public Access: 333–49. doi:10.1007/s11060-009-9877-5.

Tonomura, Noriko, Ingegerd Elvers, Rachael Thomas, Kate Megquier, Jason Turner-Maier, Cedric Howald, Aaron L Sarver, et al. 2015. 'Genome-Wide Association Study Identifies Shared Risk Loci Common to Two Malignancies in Golden Retrievers.' *PLoS Genetics* 11 (2). Public Library of Science: e1004922. doi:10.1371/journal.pgen.1004922.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. 'Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.' *Nature Protocols* 7: 562–78. doi:10.1038/nprot.2012.016.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. 'Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.' *Nature Biotechnology* 28 (5): 511–15.

Tripathi, Vidisha, Jonathan D. Ellis, Zhen Shen, David Y. Song, Qun Pan, Andrew T. Watt, Susan M. Freier, et al. 2010. 'The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation.' *Molecular Cell* 39 (6). NIH Public Access: 925–38. doi:10.1016/j.molcel.2010.08.011.

Tsai, Huai Kuang, Pei Ying Huang, Cheng Yan Kao, and Daryi Wang. 2009. 'Co-Expression of Neighboring Genes in the Zebrafish (Danio Rerio) Genome'. *International Journal of Molecular Sciences* 10 (8). Multidisciplinary Digital Publishing Institute (MDPI): 3658–70. doi:10.3390/ijms10083658.

Turri-Zanoni, Mario, Daniela Medicina, Davide Lombardi, Marco Ungari, Piera Balzarini, Cristina Rossini, Wilma Pellegrini, et al. 2013. 'Sinonasal Mucosal Melanoma: Molecular Profile and Therapeutic Implications from a Series of 32 Cases.' *Head & Neck* 35 (8). Wiley-Blackwell: 1066–77. doi:10.1002/hed.23079.

Ulitsky, Igor. 2016. 'Evolution to the Rescue: Using Comparative Genomics to Understand Long Non-Coding RNAs.' *Nature Reviews. Genetics* 17 (10). Nature Publishing Group: 601–14. doi:10.1038/nrg.2016.85.

Ulitsky, Igor, Alena Shkumatava, Calvin H. Jan, Hazel Sive, and David P. Bartel. 2011. 'Conserved Function of LincRNAs in Vertebrate Embryonic Development despite Rapid

Sequence Evolution.' *Cell* 147 (7). NIH Public Access: 1537–50. doi:10.1016/j.cell.2011.11.055.

Ulvé, Ronan, Mélanie Rault, Mathieu Bahin, Laetitia Lagoutte, Jérôme Abadie, Clotilde De Brito, Jean-Michel Coindre, et al. 2017. 'Discovery of Human-Similar Gene Fusions in Canine Cancers.' *Cancer Research* 77 (21). American Association for Cancer Research: 5721–27. doi:10.1158/0008-5472.CAN-16-2691.

Van't Hof, Arjen E, Pascal Campagne, Daniel J. Rigden, Carl J. Yung, Jessica Lingley, Michael A. Quail, Neil Hall, Alistair C. Darby, and Ilik J. Saccheri. 2016. 'The Industrial Melanism Mutation in British Peppered Moths Is a Transposable Element.' *Nature* 534 (7605). Nature Publishing Group: 102–5. doi:10.1038/nature17951.

Vassetzky, Nikita S, and Dmitri A Kramerov. 2002. 'CAN--a Pan-Carnivore SINE Family.' *Mammalian Genome : Official Journal of the International Mammalian Genome Society* 13 (1): 50–57.

Venter, J C, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, et al. 2001. 'The Sequence of the Human Genome.' *Science (New York, N.Y.)* 291 (5507). American Association for the Advancement of Science: 1304–51. doi:10.1126/science.1058040.

Vilboux, Thierry, Gilles Chaudieu, Patricia Jeannin, Delphine Delattre, Benoit Hedan, Catherine Bourgain, Guillaume Queney, Francis Galibert, Anne Thomas, and Catherine André. 2008. 'Progressive Retinal Atrophy in the Border Collie: A New XLPRA.' *BMC Veterinary Research* 4 (March). BioMed Central: 10. doi:10.1186/1746-6148-4-10.

Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F Rayner, Margus Lukk, Miguel Pignatelli, Thomas J Park, et al. 2015. 'Enhancer Evolution across 20 Mammalian Species'. *Cell*. doi:10.1016/j.cell.2015.01.006.

Villarreal, Luis P. 2016. 'Viruses and the Placenta: The Essential Virus First View.' *APMIS : Acta Pathologica, Microbiologica, et Immunologica Scandinavica* 124 (1–2): 20–30. doi:10.1111/apm.12485.

Walters-Conte, Kathryn B, Diana L E Johnson, Marc W Allard, and Jill Pecon-Slattery. 2011. 'Carnivore-Specific SINEs (Can-SINEs): Distribution, Evolution, and Genomic Impact.' *The Journal of Heredity* 102 Suppl (Suppl 1). Oxford University Press: S2-10.

doi:10.1093/jhered/esr051.

Wang, Jing, Suhas Vasaikar, Zhiao Shi, Michael Greer, and Bing Zhang. 2017. 'WebGestalt 2017: A More Comprehensive, Powerful, Flexible and Interactive Gene Set Enrichment Analysis Toolkit.' *Nucleic Acids Research* 45 (W1): W130–37. doi:10.1093/nar/gkx356.

Wang, Liguo, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. 2013. 'CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model.' *Nucleic Acids Research* 41 (6). Oxford University Press: e74. doi:10.1093/nar/gkt006.

Wang, Wei, and Ewen F. Kirkness. 2005. 'Short Interspersed Elements (SINEs) Are a Major Source of Canine Genomic Diversity.' *Genome Research* 15 (12): 1798–1808. doi:10.1101/gr.3765505.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. 'RNA-Seq: A Revolutionary Tool for Transcriptomics.' *Nature Reviews. Genetics* 10 (1): 57–63. doi:10.1038/nrg2484.

Washietl, Stefan, Manolis Kellis, and Manuel Garber. 2014. 'Evolutionary Dynamics and Tissue Specificity of Human Long Noncoding RNAs in Six Mammals.' *Genome Research* 24 (4). Cold Spring Harbor Laboratory Press: 616–28. doi:10.1101/gr.165035.113.

Watson, J D, and F H Crick. 1953. 'Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid.' *Nature* 171 (4356): 737–38. doi:10.2478/v10059-011-0025-7.

Weber, Jessica A, David H Baxter, Shile Zhang, David Y Huang, Kuo How Huang, Ming Jen Lee, David J Galas, and Kai Wang. 2010. 'The MicroRNA Spectrum in 12 Body Fluids.' *Clinical Chemistry* 56 (11). NIH Public Access: 1733–41. doi:10.1373/clinchem.2010.147405.

Wucher, Valentin, Fabrice Legeai, Benoît Hédan, Guillaume Rizk, Lætitia Lagoutte, Tosso Leeb, Vidhya Jagannathan, et al. 2017. 'FEELnc: A Tool for Long Non-Coding RNA Annotation and Its Application to the Dog Transcriptome.' *Nucleic Acids Research* 45 (8): e57. doi:10.1093/nar/gkw1306.

Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, et al. 2005. 'Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification.' *Bioinformatics (Oxford,*

England) 21 (5): 650–59. doi:10.1093/bioinformatics/bti042.

Yates, Andrew, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, et al. 2016. 'Ensembl 2016.' *Nucleic Acids Research* 44 (D1). Oxford University Press: D710-6. doi:10.1093/nar/gkv1157.

Yekutieli, Daniel, and Yoav Benjamini. 2001. 'The Control of the False Discovery Rate in Multiple Testing under Dependency'. *The Annals of Statistics* 29 (4). Institute of Mathematical Statistics: 1165–88. doi:10.1214/aos/1013699998.

Yoon, Je-Hyun, Kotb Abdelmohsen, Subramanya Srikantan, Xiaoling Yang, Jennifer L. Martindale, Supriyo De, Maite Huarte, Ming Zhan, Kevin G. Becker, and Myriam Gorospe. 2012. 'LincRNA-P21 Suppresses Target MRNA Translation'. *Molecular Cell* 47 (4). Cell Press: 648–55. doi:10.1016/J.MOLCEL.2012.06.027.

Yu, Xin, Heyi Zheng, Matthew T V Chan, and William Ka Kei Wu. 2017. 'HULC: An Oncogenic Long Non-Coding RNA in Human Cancer.' *Journal of Cellular and Molecular Medicine* 21 (2). Wiley-Blackwell: 410–17. doi:10.1111/jcmm.12956.

Zhang, Gang, Shuwei Li, Jiafei Lu, Yuqiu Ge, Qiaoyan Wang, Gaoxiang Ma, Qinghong Zhao, et al. 2018. 'LncRNA MT1JP Functions as a CeRNA in Regulating FBXW7 through Competitively Binding to MiR-92a-3p in Gastric Cancer.' *Molecular Cancer* 17 (1). BioMed Central: 87. doi:10.1186/s12943-018-0829-6.

Zhang, Kunshan, Kefei Huang, Yuping Luo, and Siguang Li. 2014. 'Identification and Functional Analysis of Long Non-Coding RNAs in Mouse Cleavage Stage Embryonic Development Based on Single Cell Transcriptome Data.' *BMC Genomics* 15 (1). BioMed Central: 845. doi:10.1186/1471-2164-15-845.

Zhou, Yunli, Ying Zhong, Yingying Wang, Xun Zhang, Dalia L Batista, Roger Gejman, Peter J Ansell, Jing Zhao, Catherine Weng, and Anne Klibanski. 2007. 'Activation of P53 by MEG3 Non-Coding RNA.' *The Journal of Biological Chemistry* 282 (34). American Society for Biochemistry and Molecular Biology: 24731–42. doi:10.1074/jbc.M702029200.

Annexes

I. Annexe 1 : Characterization and functional predictions of canine long non-coding RNAs, Le Béguec *et al*., 2018 publié dans Scientific Reports.

II. Annexe 2 : FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome, Wucher *et al.*, 2017 publié dans Nucleic Acids Research.

SCIENTIFIC REPORTS

OPEN

Characterisation and functional predictions of canine long non-coding RNAs

Received: 27 April 2018
Accepted: 24 August 2018
Published online: 07 September 2018

Céline Le Béguec[1], Valentin Wucher[1,2,3], Lætitia Lagoutte[1,4], Edouard Cadieu[1], Nadine Botherel[1], Benoît Hédan[1], Clotilde De Brito[1], Anne-Sophie Guillory[1], Catherine André[1], Thomas Derrien[1] & Christophe Hitte[1]

Long non-coding RNAs (lncRNAs) are a family of heterogeneous RNAs that play major roles in multiple biological processes. We recently identified an extended repertoire of more than 10,000 lncRNAs of the domestic dog however, predicting their biological functionality remains challenging. In this study, we have characterised the expression profiles of 10,444 canine lncRNAs in 26 distinct tissue types, representing various anatomical systems. We showed that lncRNA expressions are mainly clustered by tissue type and we highlighted that 44% of canine lncRNAs are expressed in a tissue-specific manner. We further demonstrated that tissue-specificity correlates with specific families of canine transposable elements. In addition, we identified more than 900 conserved dog-human lncRNAs for which we show their overall reproducible expression patterns between dog and human through comparative transcriptomics. Finally, co-expression analyses of lncRNA and neighbouring protein-coding genes identified more than 3,400 canine lncRNAs, suggesting that functional roles of these lncRNAs act as regulatory elements. Altogether, this genomic and transcriptomic integrative study of lncRNAs constitutes a major resource to investigate genotype to phenotype relationships and biomedical research in the dog species.

With the advancement of high-throughput sequencing technologies, transcriptome analyses (RNA-seq) have made it possible to identify major RNA classes, including the long non-coding RNA class (lncRNA)[1,2]. The transcriptome thus corresponds to sets of transcribed RNA molecules, with or without the ability to code for proteins, for a given time, condition or tissue[3]. Arbitrarily defined according to a size criterion (generally greater than 200 nucleotides), lncRNAs possess similar characteristics than RNAs encoding proteins (mRNAs), i.e. they can be spliced and have (or not) a polyadenylation tail, but they are differentiated by a lack of a functional open reading frame. Following the sequencing of all the RNA transcripts, the annotation and classification of the different RNAs consist of reconstructed transcript models, from which it is crucial to define their functional roles. LncRNAs can be located either in intergenic regions (*lincRNAs*), often close to protein-coding genes, or overlapping the opposite strand of mRNAs (*antisense*). Antisense transcription has been showed to ease a more fine-tuned regulation of gene expression, usually of the mRNA located in sense to the antisense element[4,5]. Both, lincRNAs and antisenses, represent strong candidate sequences that modulate transcription of nearby protein-coding genes.

The domestic dog has emerged as a relevant model for studying the genetic basis of numerous traits, including Mendelian and complex diseases, morphology, physiology and behaviour[6]. Genomic resources have expanded our understanding of the canine genome but exhaustive annotation of functional elements, including long non-coding RNAs, remains necessary to facilitate the identification of genotype-phenotype relationships[7]. Most recently, the number and types of canine lncRNAs have increased considerably[8,9], however, little is known about their function and biological roles. To our knowledge, a few canine lncRNAs have been characterised experimentally, as illustrated by the lincRNA close to the *BAIAP2* gene linked to podocyte migration[8] and the lncRNA *GDNF-AS* involved in a Hereditary Sensory Autonomic Neuropathy (HSAN) in hunting dogs[10]. As in other species, it is expected that lncRNAs in the dog genome are expressed at a lower level and display higher

[1]Univ Rennes, CNRS, IGDR (Institut de génétique et développement de Rennes) - UMR 6290, F-35000, Rennes, France. [2]Present address: Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona, 08003, Spain. [3]Present address: Universitat Pompeu Fabra (UPF), Barcelona, Spain. [4]Present address: UMR PEGASE, Agrocampus Ouest, INRA, 35042, Rennes, France. Correspondence and requests for materials should be addressed to T.D. (email: tderrien@univ-rennes1.fr) or C.H. (email: hitte@univ-rennes1.fr)

tissue-specificity than protein-coding genes. Also, because lncRNAs are subjected to rapid sequence turnover during evolutionary processes, they are seldom conserved in vertebrates[11]. In addition, lncRNAs are also known to regulate the expression of protein-coding genes through *cis*-acting or *trans*-acting regulation mechanisms. LncRNAs located near to and overlapping protein-coding genes have attracted particular attention from researchers and several studies in humans and mice have uncovered several mechanisms, including *cis*-regulation of the expression of their protein-coding neighbour and overlapping partners[12].

Knowledge of the biological functions of lncRNAs is continuously increasing, but their origin and evolution are still poorly understood. One recent hypothesis that has emerged concerns the theory that transposable element (TE) sequences are a possible source of non-coding exons[13,14]. In humans and mice, TEs are frequently found in lncRNAs[15] and numerous studies emphasise the importance of TEs in the regulation of gene expression (for a recent review see Chuong *et al.*[16]). In dogs, TEs also occupy a substantial fraction (40%) of the genome and might be important actors in the origin of functional novelties. More particularly, a well-studied family of TE is the specific TEs (SINEC_Cf)[17,18], that have been implicated in many dog genetic diseases[19,20] or phenotypic differences between dog breeds[21,22].

In this study, we performed an exhaustive analysis of 10,444 canine lncRNA genes for which we describe the overall transcriptional profiles across 26 distinct tissue types, representing various anatomical systems. Analysis of their expression patterns retrieved relevant relationship between tissue types and showed evidence of tissue-specificity for a large fraction of lncRNAs. Co-expression analysis of lncRNAs and nearby protein-coding genes was performed to infer putative functionality of uncharacterised lncRNAs using the principle of 'guilt by association'[23] from their co-expression with genes of known functions. Together, we provide a large and unique resource that characterises the transcriptional patterns of lncRNAs, improves the knowledge of the origins of canine lncRNAs and contributes to infer their potential functionality.

Results

Landscape of canine lncRNA transcription. We produced 16 RNA-seq in a previous study[9] consolidated by 10 RNA-seq published data[8], corresponding to a total of 26 stranded RNA-seq data to represent a panel of diverse canine anatomical systems (Supplementary Table S1). We applied a state-of-the-art RNA-seq bioinformatics pipeline[24] to assess gene expression levels for both canine lncRNAs and mRNAs (see the Methods section). For lncRNAs, we used the FEELnc-based classification[9] based on the international gold-standard lncRNA classification given by the GENCODE consortium[25] to differentiate long intergenic non-coding RNAs (lincRNAs, n = 5,651) from antisense lncRNAs that have overlapping protein-coding genes in antisense (n = 4,793). We determined that a total of 7,763 (74.3%) lncRNAs are expressed in at least one tissue with a normalised count[26] TPM > 1 where TPM = Transcripts Per Million and 9,542 (91.4%) with a TPM > 0.1. By comparing their respective expression levels with protein-coding genes mRNAs, canine lncRNAs have, on average, 20 times lower expression levels (Wilcoxon test, p-value $< 2.2^{e-16}$) for all tissues, with the notable exception of the testis tissue for which the difference is 6 times less pronounced, a trend also showed in other species[2,27] (Fig. 1a). We then compared the expression of lncRNAs in various tissue types and found that, as expected, the mean expression level in testis tissue is much higher (TPM = 7.02) than in any other tissue (ranging from 0.11 for the pancreas tissue to 1.60 for the hair follicle tissue and with an average TPM = 1.05) (Fig. 1a). By comparing lincRNA versus antisense expression patterns (Supplementary Fig. S1), we observed that lincRNAs have a lower expression than antisense, except for testis tissue (Wilcoxon test, p-value $< 2.2^{e-16}$).

Taking advantage of the variety of samples, we analysed whether canine lncRNAs displayed patterns of tissue-specific expressions by calculating a tissue-specific score 'tau'[28]. This score, ranging from 0 to 1 (with tau = 0 being broadly expressed and tau = 1 being a highly tissue-specific expression), has been shown to be highly robust when evaluating tissue-specific genes[29]. Using a stringent threshold ≥0.95, a total of 44% of lncRNAs (n = 4,599, Supplementary Table S2) displayed a clear pattern of tissue-specificity, highlighting the potential specialised functionality and distinct spatial pattern of expression in the corresponding tissue. In comparison, only 17% of mRNAs (n = 3,635) showed a pattern of tissue-specificity reflecting that, as in other species, canine lncRNAs are more tissue-specific than mRNAs[30]. Among all classes of lncRNAs, we observed that lincRNAs are more tissue-specific (68.6%) when compared to antisense (31.4%) (Chi-square test, p-value $< 2.2^{e-16}$). As shown in humans, the canine testis tissue is particularly enriched in tissue-specific lncRNAs[1,2] (n = 3,001). This highlights the singularity of this tissue, probably due to the presence of many cell types, the state of its chromatin and the binding of specific transcription factors[31]. Excluding testis tissue, we identified an average of 63 lncRNAs expressed specifically by tissue. In order to assess the relative enrichment of tissue-specific lncRNAs versus tissue-specific mRNAs, we measured the proportion of tissue-specific genes amongst the total number of expressed genes (Fig. 1b). As an example, this highlighted that 20% of lncRNAs expressed in liver tissue are liver-specific, compared to only 2% of mRNAs.

Given that tissue-specificity can be influenced by cells with common origins, we then investigated the tissue-specificity within tissues related to the nervous system including brain, cerebellum, cortex, olfactory bulb and spinal cord. We found specific expression patterns for 260 lncRNAs in the different samples of the nervous system[32]. Among the nervous system samples, we identified 156 lncRNAs specifically expressed in the cerebellum, 60 in the cortex and 44 in the olfactory-bulb (Supplementary Table S2); (tau threshold ≥0.95). These lncRNAs were characterised by a higher expression level (mean TPM = 3.30) in comparison to other tissues (mean TPM = 0.12). For example, we found a cerebellum-specific lincRNA (tau = 0.96), annotated in humans as *CASC18*, with a role in neural cell differentiation[33].

Although we identified a large number of tissue-specific lncRNAs, we then clustered the 26 tissue samples based on their lncRNA expression profiles in order to detect groups of lncRNAs that exhibit common expression patterns. The heatmap in Fig. 1c highlights that clustering the lncRNAs' expression data recovers biologically meaningful relationships between tissue types. This analysis defined two main clusters with more than 3 samples,

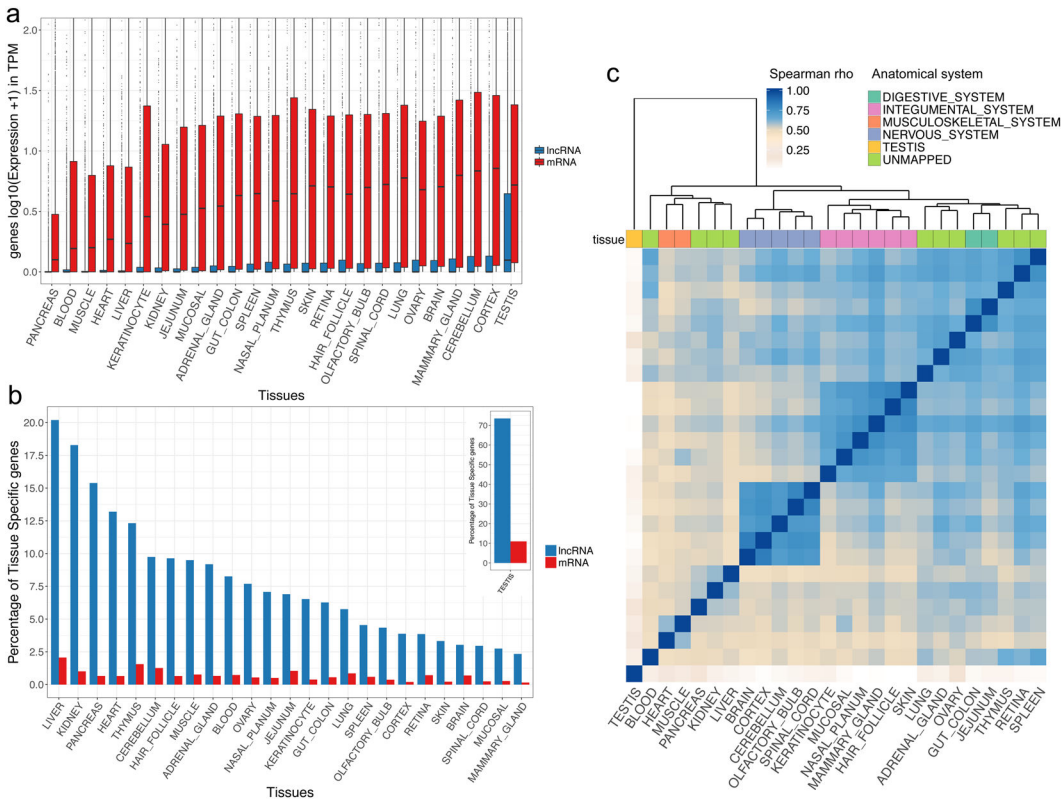

Figure 1. (**a**) Comparative analysis of log10 transcription (TPM + 1) levels between mRNAs (red) and lncRNAs (blue) genes in 26 canine tissues. (**b**) Proportions of tissue-specific lncRNAs (blue) and mRNAs (red) amongst expressed genes in each tissue with testis tissue represented in the separated panel (top right). (**c**) Hierarchical clustering of 26 tissues based on Spearman correlations measured from lncRNA expression data.

which grouped nervous system tissues (n = 5) and integumental system tissues (n = 6) and supported the fact that some lncRNAs shared expression patterns in multiple tissues. This analysis was repeated using only mRNA expressions and showed similar clusters (Supplementary Fig. S2). Complementary to the previous tissue-specific analysis, this clustering approach allowed us to identify an lncRNA (*RLOC_00033166*) expressed in all nervous system tissues with a mean TPM = 7.89 but not detected in any other tissues. Interestingly, this lncRNA is transcribed in antisense orientation to the Neuregulin 3 gene (*NRG3*), which is involved in neuroblast cell differentiation[34] and thus represents a potential candidate for its regulation. Another lncRNA (*RLOC_00020746*) is almost only expressed in the musculoskeletal system cluster (heart and muscle) and is expressed in antisense to the Popeye Domain Containing 3 gene (*POPDC3*), which may play an important role in the development of skeletal muscle and heart tissues[35]. Apart from the clustered groups of tissues identified in the analysis, most samples, and particularly the testis sample, remained strongly unrelated to any other tissues.

Tissue-specific expression of lncRNAs correlates with their transposable element content. TEs are thought to constitute part of the active lncRNA regions[13] and may drive the specific expression patterns observed for lncRNAs[16]. Interestingly, in dogs, a canine-specific SINE repeat family (SINEC_Cf) has been shown to provide many polymorphic site insertions[17] and is also associated with multiple disease phenotypes[16,18]. To gain insights into the relationships between canine lncRNAs and transposable elements, we analysed the TE content of the 10,444 lncRNAs in the 4 major TE classes (DNA transposons, LTRs, LINEs and SINEs) as annotated by the RepeatMasker database[36]. We first showed that almost 70% of canine lncRNA transcripts contain at least one TE overlapping an lncRNA exon, a result in between human (83%) and mouse (66%) lncRNAs[15]. We then determined that ~20% of the cumulative exon sequences of lncRNAs are composed exclusively of TEs (Fig. 2a).

This proportion of lncRNA exonic TEs is less than that observed in the entire dog genome (37.7%) but is 2.5 times higher than for mRNAs exonic sequences (7.9%), highlighting the high prevalence of TEs in dog lncRNA exons as also observed in other species[15]. Compared to the genome-wide distribution of TEs, most TE families are under-represented in lncRNAs with the exception of LTR-ERVL retroviruses (Long Terminal Repeat-Endogenous RetroVirus-Like elements) (Supplementary Fig. S3), which are also significantly enriched in human lncRNAs[15]. In addition, we computed canonical correlations between TE content in both lncRNAs exons and promoters and lncRNA tissue expressions using the mixOmics program[37] (see Methods). This analysis highlighted that lncRNA expression in the different tissues tends to be associated with high contents of SINEC_Cf specifically in lncRNA promoters (Fig. 2b), a pattern not seen for mRNAs (Supplementary Fig. S4). These results suggest that changes in the promoter sequences of lncRNAs by TE insertions and more particularly SINEC_Cf, might contribute to

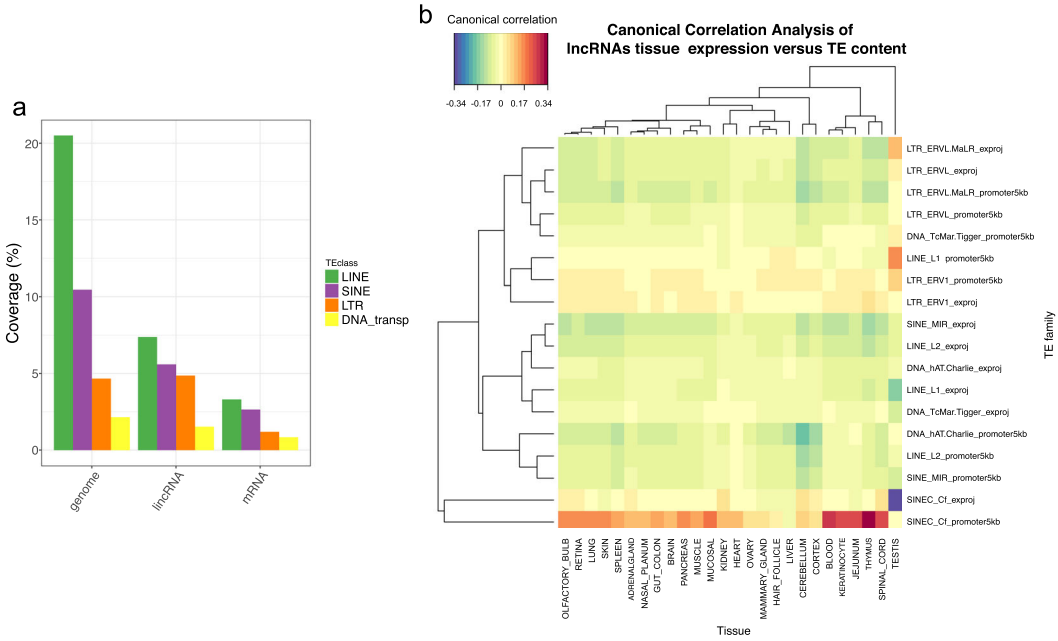

Figure 2. (**a**) Proportion of transposable elements (LINEs, SINEs, LTRs and DNA transposon in green, purple, orange and yellow respectively) with respect to the entire dog genome, lncRNA exons and mRNA exons (from left to right). (**b**) Canonical correlations between lncRNAs' expression in the 26 tissues (column) and TE families in exons and promoters (row).

the specific spatial expression of canine lncRNAs. The analysis also pinpoints that exons of lncRNAs expressed in testis are particularly depleted in SINEC_Cf, suggesting that testis-specific expression of canine lncRNAs is more likely associated with other family of TEs such as LTR (Fig. 2b) and probably LTR-ERVL (Supplementary Fig. S3) as recently highlighted in mouse germline[38].

Identification of dog-human conserved lncRNAs by comparative genomics. DNA conservation among divergent species is a widely used indicator to infer functionality and to suggest conserved functions. Identifying orthologous lncRNA relationships is challenging since lncRNAs are more likely to be gained or lost during evolution[11] than constrained sequences as protein-coding genes. Therefore, we used a positional comparative genomics approach, parameterised at the synteny level, to predict putative lncRNA orthologues between dog and human using the EnsEMBL Compara database[39] (see Methods). Using this strategy, we mapped the positions of canine lncRNAs onto the human genome and identified 939 lncRNAs (9.0%) with 1:1 human non-coding orthologues from the GENCODE database[25] which can be separated into 560 lincRNAs (9.9% of the total number of lincRNAs in dogs) and 379 antisense (7.9% of the total number of antisense in dogs) (Figs 3 and S5). We found that lncRNA orthologues that were part of a synteny block shared similar gene structures, with a mean number of exons per gene of 3.6 and 3.1 for humans and dogs, respectively (Supplementary Table S3). Based on the GENCODE annotation of the conserved dog-human lncRNAs, we were able to confirm the annotation of the well-studied *HOTAIRM*[40], *MALAT*[41], *NEAT_1*[42] and *PCA3*[43] in dogs and to infer novel canine orthologous relationships such as the CASC family of lncRNAs (*CASC15*, *CASC17*, *CASC18* and *CASC20*) and *INHBA-AS1*[44] or *MEG9*[45]. Complementary to this synteny-based strategy, we evaluated the selective constraints acting on canine lncRNAs and mRNAs by using the Genomic Evolutionary Rate Profiling (GERP) scores[46] (see Methods). In comparison to mRNAs, we showed that the 939 syntenically conserved lncRNAs are significantly less constrained than syntenic mRNAs (Wilcoxon test, p-value $= 9.54^{e-171}$). In addition, this allowed us to pinpoint that syntenically conserved lncRNAs tend to be more constrained (mean GERP = 0.221) than non-syntenic lncRNAs (median GERP = 0.195), although not statistically significant (Wilcoxon test, one-tailed p-value = 0.19). Altogether, this is in agreement with other studies that have reported a lower purifying selection acting on lncRNAs[47].

Analysis of transcriptional profiles of conserved lncRNAs through comparative transcriptomics. As sequence conservation is not an infallible indicator of functionality, we integrated the expression patterns of orthologous genes in our analysis. It has now been reported that orthologous genes may have conserved or variable expression profiles between species such as in human and mouse[48]. To investigate the expression patterns of orthologous lncRNAs between dogs and humans, we determined the correlation of the level of expression between 11 human tissues from the ENCODE project[49] that matched the dog tissues. We produced bar charts (Fig. 4) to represent the expression levels of each conserved lncRNA in the 11 different samples of both species and these allowed us to visualise and compare expression patterns of orthologous lncRNAs between dogs and humans (Supplementary Fig. S6). Overall, of the 727 conserved lncRNAs with expression in at least one tissue on both species, mean lncRNA expression patterns correlate well between dog and human meaning that when

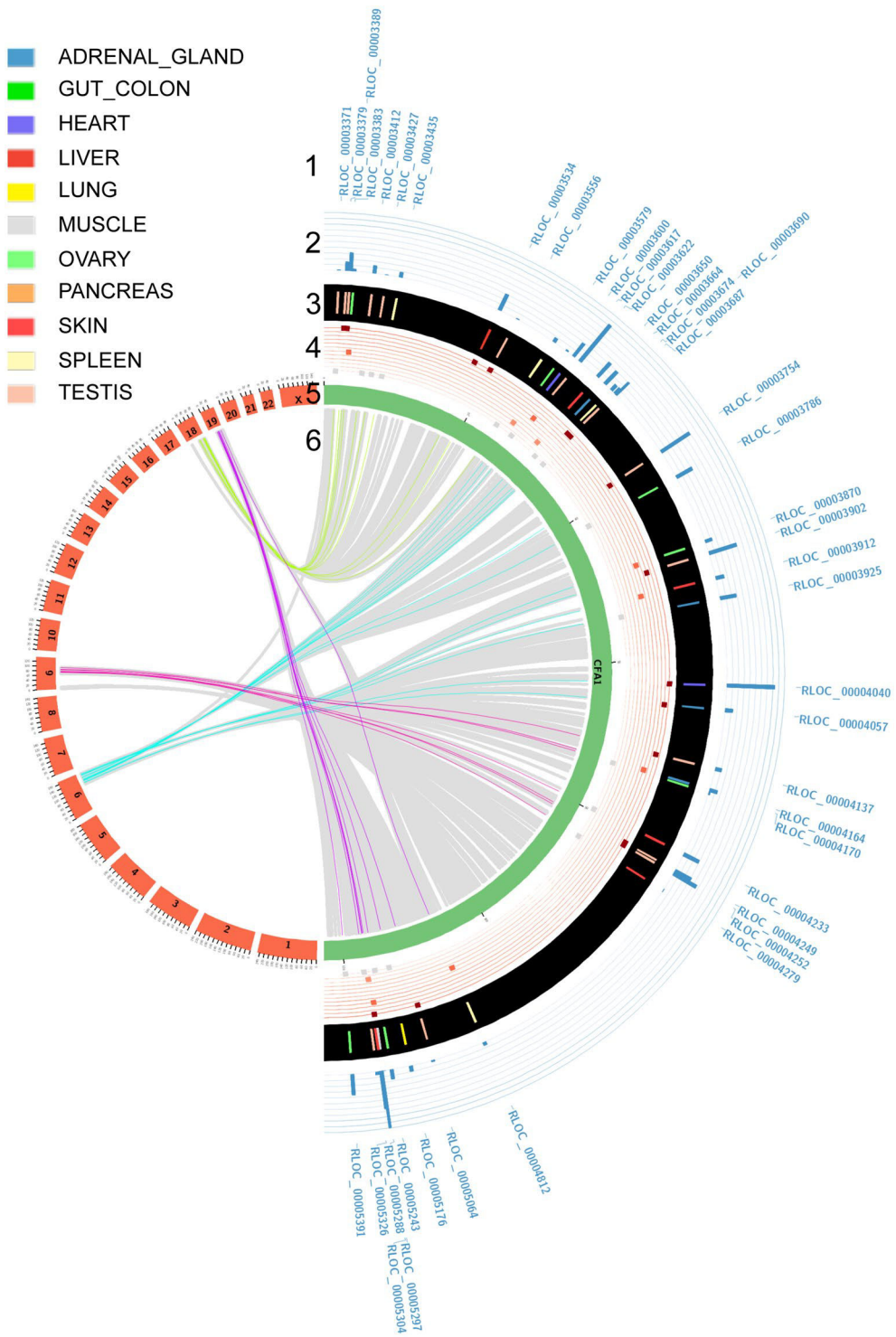

Figure 3. The Circos plot provides the visualisation of dog-human orthologous relationships of lncRNAs and their comparative transcriptomic-based expression patterns. Here, we provide plot for canine chromosome 1 the largest chromosome which allows to summarize the most amount of information. Tracks are described from the outside to the inside. Track1: Labels of canine lncRNAs identified with human orthologues are described at the most outside track of the figure. Track2: Level of expression of canine lncRNAs is shown by the blue histogram. Track3: The tissue in which lncRNA is expressed in dogs. The 11 tissues are adrenal gland, gut-colon, heart, liver, lung, muscle, ovary, pancreas, skin, spleen and testis and are represented with the colour code as shown in the legend. Track4: Expression of the human orthologous lncRNA is represented. 11 lines are reported. When the human orthologue lncRNA is expressed in the same tissue and with the highest of expression, it is represented by a dark red square on the upper line. When an orthologous human lncRNA is expressed in the same tissue, with the second maximum of abundance, it is represented on the second line in lighter red. When an orthologous human lncRNA is not expressed in the same tissue, it is represented by a grey square. Track5:

The green layout depicts the canine chromosome, the red layout represents the human chromosomes. Track6: For the innermost part, coloured lines link the dog-human orthologous relationships of lncRNAs. Grey lines represent the orthologous relationships of protein-coding genes.

a canine lncRNA is expressed, its human orthologue is also transcribed in the same human tissues with the same orders of magnitude (mean correlation rho = 0.39). Among these general tendencies, we could identify a subset of 26% of dog-human orthologous lncRNAs (n = 192) with expression patterns significantly correlated (p-value < 0.05; mean rho = 0.87), suggesting that these lncRNAs might be involved in evolutionary conserved functions. In addition, these lncRNAs are highly tissue-specific (average tau = 0.94 and 0.93 for dogs and humans, respectively), a finding already reported between human and mouse where tissue-specific lncRNAs in one species are also tissue-specific in other[48]. These evolutionary and transcriptionally conserved lncRNAs are mostly expressed in testis (n = 154) then spleen (13), lung (9), liver (8), skin (5), and heart (4) (Fig. 4a). Anatomical systems such as the brain, testis, heart, liver and kidney were previously found to have clearly distinct signatures of tissue-specific genes in humans and mice[48] and showed strong conservation between the two species. Here, as an example, the lncRNA *TRDN-AS1*, transcribed in the opposite strand of the Triadin gene (*TRDN*), is found specifically expressed in the heart tissue in both human and dog species. Another example is given by the *LINC01698* gene which exhibits a shared expression pattern between dog and human in the skin, as described using GTEx data portal[50]. We also defined lncRNAs (n = 77) with intermediate correlation pattern (p-value < 0.2; mean rho = 0.50), which display a more tissue-variable expression between species (Fig. 4b). This subset shows that sequence constraints do not imply an exact conserved expression and that different factors in the two species may act on the expression of the same gene, modulating its expression level. A third subset of lncRNAs (n = 458) was defined with divergent transcriptional pattern (p-value > 0.2; mean rho = 0.17) between the two species (Fig. 4c). This subset may indicates expressed lncRNAs in anatomical systems with complex structures composed of many primary cell types, for which tissue functionality may differ between samples and/or between species because of their substructure.

LncRNA:mRNA co-expression and correlation analysis. In many species, lncRNAs have been involved in regulating gene expression by coordinating epigenetic, transcriptional or post-transcriptional processes[51]. Co-expression pattern analyses provide a means to investigate sequences that modulate the transcription of nearby genes. In addition, regulatory sequences are often located near to their target gene, modulating their transcription. Hence, identifying correlated expressed sets of protein-coding and non-coding loci provides a way to (i) identify relationships between lncRNAs and mRNAs and (ii) infer potential regulatory functionality of lncRNAs.

Here, we searched for significant co-expression, computed through statistical correlations between lncRNAs and their neighbouring protein-coding genes. We used the 'guilt-by-association'[23] strategy to functionally annotate the lncRNA co-expressed with its mRNA neighbour, especially in the case of divergent transcripts sharing bidirectional promoters[52]. We retained neighbouring protein-coding genes transcribed in divergent and convergent orientations (excluding same strand) with respect to the lncRNA in order not to bias the correlation analysis with lncRNA being actually unannotated UTRs of the neighbour protein-coding genes. On average, these canine lncRNAs are located at 25 kb of the closest protein-coding genes. We identified more than 126,000 lncRNA:mRNA pairs within a 1 Mb window using the FEELnc classifier module[9]. We then divided pairs into two classes consisting of antisense:mRNA pairs, defined by lncRNAs overlapping mRNAs transcribed in antisense (n = 3,401) and lincRNA:mRNA pairs defined by intergenic lncRNAs located at less than 1 Mb of an mRNA (n = 123,456). Using the expression data from the 26 tissues, we computed expression correlations and mined the resulting pairs. We found 8,139 significant correlations (| rho | > 0.5 and p.adjust BH[53] < 0.05) consisting of 7,615 lincRNA:mRNA and 524 antisense:mRNA pairs for a total of 3,410 distinct lncRNAs. These results revealed co-expressed pairs that may predict regulatory relationships. Among significant correlations, only 58 pairs were found with a significant negative correlation coefficient (rho < −0.5). The lncRNA co-expressed in these pairs are significantly more distant to their mRNAs (mean distance = 471 kb) compared to the ones in the positive correlations (mean distance = 394 kb) (Wilcoxon test, one-tailed p-value = 0.02), indicating that they may be located in different genomic regions. An interesting example of positive correlation is given by the lincRNA *RLOC_00018074* co-expressed with the mRNA *EGFR* (Epidermal Growth Factor Reception) in the 26 tissues (Fig. 5a). *RLOC_00018074* and *EGFR* genes are transcribed in a divergent orientation and, by using published data which mapped dog promoters from H3K4me3 marks[54] (Fig. 5b), we could hypothesise that this pair share a bi-directional promoter thus reinforcing the validity of our functional inference.

Of the 3,410 distinct lncRNAs, we observed multiples type of co-expression patterns. A subset of 1,691 (49.6%) lncRNAs have two or more mRNAs significantly co-expressed for which more than 800 (55%) displayed a higher correlation with the second nearest protein-coding gene or greater (Supplementary Fig. S7). It shows the importance of identifying co-expression patterns for establishing accurate lncRNA:mRNA pairs that may interact. In addition, the strength of the correlation decreases with the distance between the two elements of the pair, as we computed with a Spearman correlation analysis (Supplementary Fig. S8). Indeed, highly co-expressed gene pairs (0.9 < | rho | ≤ 1) are closer (average = 89 kb) than less correlated co-expressed gene pairs (0.5 < | rho | < 0.7; average = 411 kb).

Finally, we performed a gene ontology (GO) term enrichment analysis in order to predict potential lncRNA functions. Based on the human orthologous mRNAs (n = 3,977) with which the lncRNAs are co-expressed, we focused our GO analysis on the Biological Process (see Method). A total of 22 GO terms were significantly enriched (adjusted p-value < 0.05) and correspond to developmental processes such as 'sensory organ

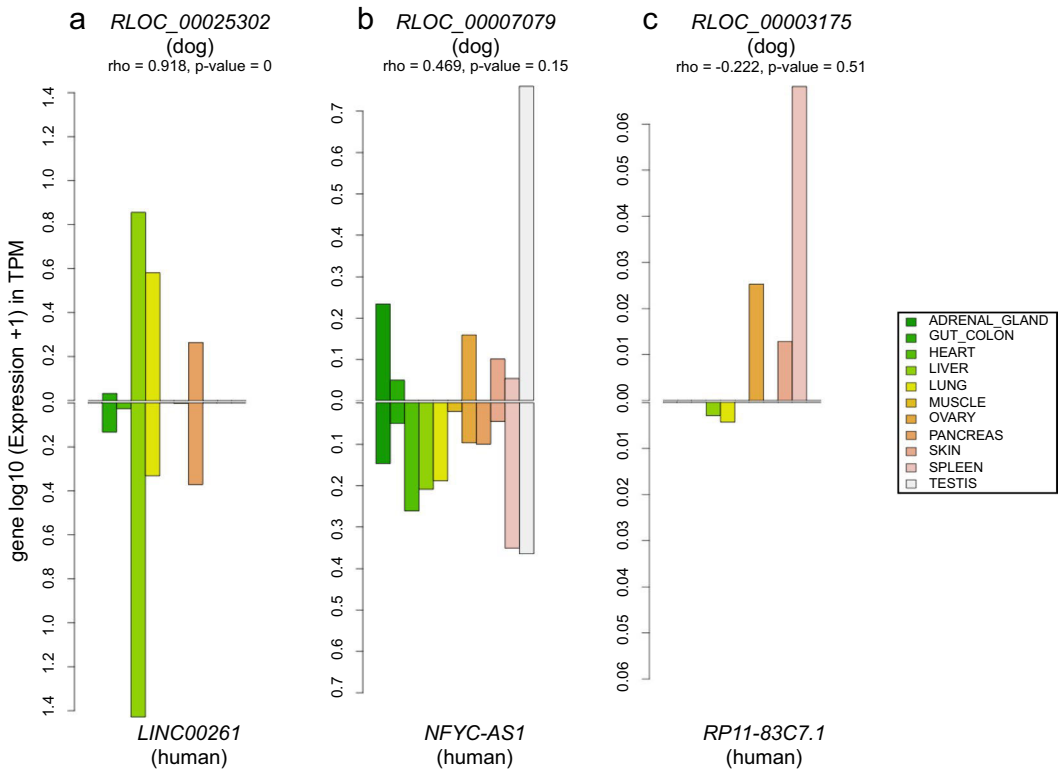

Figure 4. Bar charts representing the expression of the 3 lncRNAs in 11 matched dog-human tissues. (**a**) High conservation of expression, (**b**) intermediate conservation of expression, (**c**) low conservation of expression.

development' (GO: 0007423, 147 genes), 'axon development' (GO: 0061564, 129 genes) or 'hindbrain development' (GO: 0030902, 49 genes) (Supplementary Table S4). These 22 GO terms included 913 unique mRNAs co-expressed with their corresponding canine lncRNA (26.8%). These results predict a functional assignment for these 913 canine lncRNAs which represent interesting candidates for studies related to the enriched biological processes.

Discussion

Transcriptomic studies have emphasised the analysis of long non-coding transcripts using expression profiles to characterise the patterns and their potential functional roles. LncRNAs tend to be expressed at lower levels than protein-coding genes, being under less stringent evolutionary constraints, and are preferentially enriched in regulatory functions. Here, we realised a global analysis of long non-coding RNA expression across multiple tissues to aid genome annotation and improve functional annotation of the dog genome. By using strand-specific RNA sequencing of 26 tissues, we profiled the expression patterns of lncRNAs and achieved a detectable expression for 91% of lncRNA in at least one tissue, which constitutes the largest resource of long non-coding RNA expression data sets in dogs. We showed the higher tissue-specificity pattern of lncRNAs relative to mRNAs, and this suggests specialised functions in the development, differentiation and physiological processes of tissues. As also observed in other species, the largest number of tissue-specific lncRNAs was shown to be in the testis, a result that can be related to the pervasive transcription during spermatogenesis process and due to chromatin remodelling[55].

There is growing evidence showing a close association of transposable elements (TEs) with non-coding RNAs. Thousands of long intergenic non-coding RNAs are associated with endogenous retrovirus LTR TEs in human cells[13]. Exhaustive characterisations of the links between lncRNAs and TEs are becoming fundamental as diseases and phenotypic traits are increasingly found to have a TE and/or lncRNA etiology[13,16]. Because of the absence of open reading frame and thus amino acids constraints, lncRNAs sequences are more tolerant to TE insertions, and some authors hypothesized that repeat sequences into lncRNAs are the functional domains of these elements[13]. In addition, numerous examples of TE insertions having constructive effects because of their inherent ability to bind to RNAPII and provide *cis*-regulatory activities[16].

Here, we have shown a strong association of the SINEC_Cf TE family with lncRNAs in dogs. The SINEC_Cf family have the capacity to spread in the genome[17] by the mean to insert additional copies of themselves and, therefore, may be able to introduce sequences with regulatory roles into lncRNAs or their promoter sequences upon insertion, and might contribute to the spatial expression of canine lncRNAs. These results provide the first evidence for discovering how and which TE families represent a major force in shaping the canine lncRNA expression and their tissue-specificity expression patterns in a lineage-specific fashion.

Despite the number of tissues investigated in this study (n = 26), tissue-specificity is a limiting factor in annotating the whole repertoire of lncRNAs of a species, since their annotation depends on the availability of

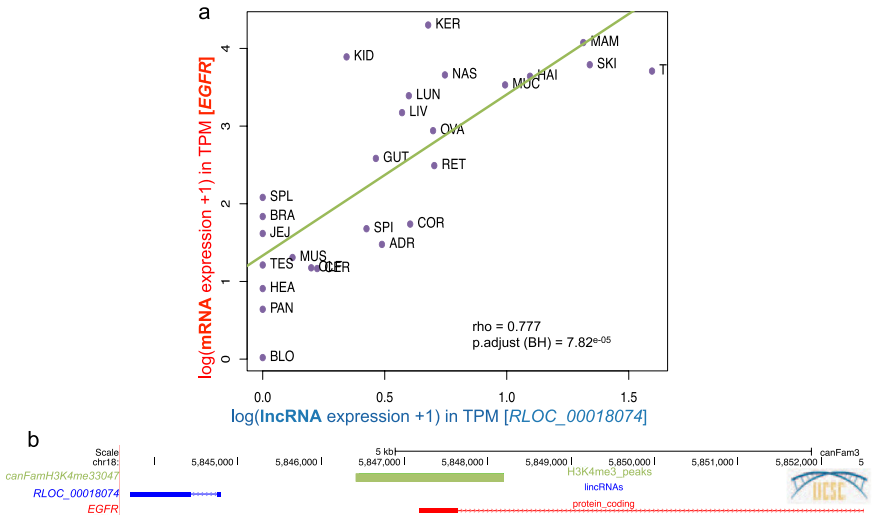

Figure 5. (**a**) Correlation plot of the lncRNA:mRNA pair (*RLOC_000018074:EGFR*) in the 26 tissues. (**b**) UCSC screenshot representing the genomic localisation of the *RLOC_000018074:EGFR* pair transcribed in a divergent orientation with the promoter mark (H3K4me3) from published data.

numerous tissue and cell types as well as tissue subtypes and multiple spatio-temporal conditions. Additional tissues and subtypes and development stages must be continuously collected through large databases as we do with the Cani-DNA French database (https://dog-genetics.genouest.org/en/brc-cani-dna/) in order to provide additional anatomical subtype data for identifying an extended catalogue of lncRNAs and their isoforms. Until now, most expression studies have been performed on a population level usually averaging the transcriptomes of millions of cells[56]. Pioneering RNA-seq of single cells has provided the characterisation of transcriptional differences of both coding and non-coding RNAs on a genome-wide scale[57]. The single-cell RNA-seq approach (scRNA-seq), a relatively recent experimental approach, in contrast to bulk RNA-seq sequencing, is a new technology, for which the main strength is the ability to measure the distribution of expression levels for each gene across a population of cells. These data make it possible to study questions for which transcriptome-specific modifications are important, such as the identification of the cell type, the heterogeneity of cellular responses, the stochasticity of gene expression, or the definition of gene regulatory networks through the cells. In particular, the use of scRNA-seq would make it possible to individually sequence all the cells making up the cycle of spermatogenesis and to identify more precisely certain phenomena that bind to one or the other of the steps. Or in the case of tissues with high heterogeneity and complexity, scRNA-seq will be able to detect and characterise the expression patterns of complex tissues, in order to provide high-resolution identification of cell types and markers and to identify splicing patterns and allelic random expressions that are variable between cells. ScRNA-seq will help in-depth, functional characterisation and gene expression variation derived from different genetic backgrounds at the cell resolution.

Co-expression analyses rely on similarities in gene expression and are widely used for the functional annotation of unknown genes using the principle of 'guilt-by-association'[23]. Focusing on neighbouring genes, we tested the expression correlation of lncRNA:protein-coding pairs and observed that thousands of lncRNA-coding pairs showed an overall positive correlation, suggesting coordinated transcription and, by implication, a shared function or pathway. Whilst constructing co-expression networks is rather straightforward, the resulting network of connected lncRNA:mRNA do not provide information on the nature of the regulatory relationship of the connected genes, which limits its biological interpretation.

Next-generation sequencing technologies have considerably advanced the field of comparative genomics and comparative transcriptomics. These approaches are particularly important for studying the evolution of gene regulation in model organisms, investigating the level of the lncRNA sequence and expression conservation with humans and investigating the way that this conservation determines conditions in which the dog constitutes an appropriate model for diseases and phenotypic traits. Here, we identified more than 900 lncRNAs with a human orthologue and the analysis of the level and distribution of expression between dog and human, despite a low number of samples, allowed us to define conserved or diverged expression patterns that then served to predict putative, conserved functions or to pinpoint more subtle changes of functionality. A previous comparative transcriptomic analysis between human and others species, was also able to define conserved or diverged expression patterns using only 6 common tissues[58]. Much comparative data still needs to be produced and analysed concerning transcriptional changes associated with differentiation and development or with cellular responses to external stimuli.

In this study, we functionally annotate more than 900 canine lncRNAs with the enriched functions among the protein-coding gene sets of distinct GO functional categories that are co-expressed with each lncRNA. While this GO analysis revealed numerous lncRNAs associated with diverse biological processes, a part of the enriched functions have been validated by published studies[59,60].

In addition, future functional studies will be important to investigate roles of lncRNAs by current approaches of loss- and gain-of-function biological experiments[61,62]. Consequently, there is a need for large-scale, systematic approaches to analyse the functional contribution of lncRNAs using genome-scale CRISPR interference (CRISPRi). In a recent study, Liu *et al.*[63] developed genome-scale CRISPRi approach which targets 16,401 lncRNA genes each with 10 sgRNAs per transcription start site for identifying functional human lncRNA.

The dog is a promising model for complex phenotypes, genetic diseases and clinical studies. Altogether, this genomic and transcriptomic integrative study of lncRNAs constitutes a major resource for the dog species.

Materials and Methods

Datasets. The 26 RNA-seq sample dataset (Supplementary Table S1) represents a wide variety of canine cell-types and tissue-types. The 26 tissues can be classified by Anatomical System as inspired by the Expression Atlas database (https://www.ebi.ac.uk/gxa/home/), as shown in the Supplementary Table S1. Two tissues correspond to 'Digestive system', 6 tissues to 'Integumental system', 2 tissues to 'Musculoskeletal system' and 5 tissues to 'Nervous system'. The others tissues correspond to 'Unmapped' and 'Testis'.

RNA extraction and directional sequencing were performed as described in Wucher *et al.*[9] and Hoeppner *et al.*[8] and are available through accession numbers SRP077559, SRX111061 - SRX111071, and SRX146606 - SRX146608.

The canine 'canFam3.1-plus' annotation[9] (containing 10,444 lncRNA and 21,810 protein-coding genes (mRNAs)) was used as the reference annotation for this study and the canFam3 assembly version was the reference genome.

Quantification of mRNAs and lncRNAs expressions. Based on the bio-informatic protocol described in Djebali *et al.*[24], FASTQ reads were aligned on the canine reference genome and transcriptome using the STAR[64] program (version 2.5.0a) and we determined the gene and isoform expression levels for both lncRNAs and mRNAs for each of the 26 RNA-seq with RSEM[26] (version1.2.25). Finally, the RSEM output files were parsed in order to extract gene expression level, normalised in Transcripts Per Million (TPM) and merged to obtain one matrix expression file with gene names in rows and tissue expression levels (TPMs) in columns.

Characterisation of lncRNA expression. *Tissue-specificity.* To calculate the tissue-specificity score 'tau'[28], ranging from 0 to 1 (with tau = 1 for highly tissue-specific genes and tau = 0 for broadly expressed genes), we used the matrix file described above and calculated the following equation[28]:

$$\tau = \frac{\sum_{i=1}^{n}(1 - \hat{X}i)}{n - 1}; \hat{X}_i = \frac{X_i}{\max_{1 \le i \le n}(x_i)}$$

where n corresponds to the number of tissues analysed and x_i is the gene expression in the tissue i. The specificity score was calculated using a threshold of 1 TPM (to be stringent).

To extract a stringent set of tissue-specific genes, we used a tau threshold greater than 0.95 corresponding to a ratio of 4 between the first and the second tissue (e.g. $\max(x_i)/\max^{2nd}(x_i) > 4$) or, in others words, corresponding to cases where gene i is four times more expressed in the first tissue compared to the second most highly expressed tissue. Note that we provide a perl script dedicated to computing the 'tau' score on GitHub available here:

https://github.com/tderrien/IGDR/blob/master/script/specificity_score.pl.

Finally, the 'tau score' together with eight others metrics dedicated to computing tissue-specificity were recently benchmarked in Kryuchkova-Mostacci *et al.*[29], and the authors concluded that "the best overall method to measure expression specificity appears to be Tau" because it consistently appears to be the most robust method for retrieving tissue-specific genes.

Heatmap and clustering analysis. For each biotype (lncRNAs and mRNAs), we computed all pair-wise Spearman correlations between the 26 tissues in order to obtain a matrix of distance. Using this file, together with a meta-data file containing anatomical system classification (e.g. nervous system, integumental system, musculoskeletal system, digestive system, testis, and unmapped), the heatmap and the associated hierarchical ascendant clustering (with default parameters: euclidean distance and method = 'complete') were created as in Breschi *et al.*[48], using R software (version 3.2).

TE content. Canine TE annotations were based on the RepeatMasker[36] database, downloaded from the UCSC table browser[65] (September 2017). Four major classes were analysed: DNA transposons and retrotransposons, LTRs, LINEs and SINEs; each of them being divided into families (e.g. SINEC_Cf). To compute the proportion of TEs per gene, exons of each gene were projected onto the genome and the number of exon-projected nucleotides overlapping TE families was computed using the bedtools program (version 2.19.0) on bed12 input files with the -wao and -split options. For gene promoters, the same process was repeated using 5 kb upstream of each gene starts.

TE content and gene expression. The mixOmics tool[37] (version 6.3.1) was used to compute canonical correlations to highlight potential correlations between two matrices X and Y, where X is the matrix of lncRNA (or mRNA) expressions in the 26 tissues and Y is the matrix of TE content in lncRNAs (or mRNAs) exons and promoters. The rcc function of the mixOmics package (with method = 'ridge' and ncomp = 3) was used to seek for linear

combinations of the variables (canonical variables) to reduce the dimension of the datasets. The heatmap was constructed with the mixOmics function cim (Clustered Image Maps) with default values, except ncomp = 3.

Comparative genomics and transcriptomics. *Comparative genomics.* To identify conserved canine lncRNAs and mRNAs in humans, we used the Ensembl Compara database[39]. Briefly, the program annotates orthologous regions between two or more species based on Whole Genome Alignments (WGA) computed using the EPO (Enredo-Pecan-Ortheus) pipeline[39]. All canine gene coordinates were then mapped onto the human genome (version: GRCh37 using version 75, GENCODE version 19), having the same biotype (coding or non-coding) and considering only the one-to-one (1:1) relationships.

GERP score. We used the EnsEMBL Compara API to retrieve GERP scores[46] of the dog genome assembly (canFam3) (one wig file per each dog chromosome) computed on the 53 eutheria mammals. To get a single value per lncRNA and mRNA gene, we projected all exons of each gene onto the dog genome and calculated the median GERP score for each gene.

Comparative transcriptomics with human ENCODE data. We extracted the human gene expression using data produced by the ENCODE project[49] (https://www.encodeproject.org/matrix/?type=Experiment&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&biosample_type=tissue&assay_title=total+RNA-seq&assembly=hg19&award.project=ENCODE). We only selected data where tissues between humans and dogs could be matched. We found 11 dog:human matching tissues: adrenal gland, gut-colon, heart, liver, lung, muscle, ovary, pancreas, skin, spleen and testis. From the available RSEM files, we extracted the expression levels at the gene transcript levels in TPM for the dog data. When multiple RSEM files for a tissue replicates, we averaged the expression level per gene. We then determined the statistical correlation, with the Spearman correlation test, of the level and the distribution of the expression between these 11 human and dog tissues. Bar charts were created with R software to help visualise the comparison.

Comparative genomics and transcriptomic visualisation. We used the Circos tool (circus-0.69-6) to visualise relationships of orthologous genes between dog and human. Circular layouts were constructed to allow global but meaningful figures with a Circos ideogram per dog chromosome.

Co-expression analysis. *Co-expression.* To study the correlations of expression between lncRNAs and mRNAs, we used the FEELnc classifier module[9] to annotate lncRNA classes and sub-classes (e.g. genic versus intergenic and divergent versus convergent, Supplementary Fig. S9) and to automatically identify all lncRNA:mRNA pairs in a sliding window of 1 Mb around each lncRNA. For the sake of interpretability, we filtered subclasses so that they only retained exonic antisense, intergenic divergent and intergenic convergent. For all lncRNA:mRNA pairs, we computed Spearman correlations between expression vectors (26 points) and corrected, associated p-values for multiple testing using the Benjamini-Hochberg (BH) method[53]. To identify significant pairs in co-expression, we used a correlation coefficient of $|\text{rho}| > 0.5$ and $\text{p.adjust} < 0.05$.

GO term analysis. From the 5,711 canine mRNAs that were significantly co-expressed with their lncRNAs (n = 3,410), we extracted human, orthologous EnsEMBL gene IDs of mRNAs (n = 3,977). To identify gene ontology (GO) terms significantly enriched in this list of gene IDs, we used the web-based GEne SeT AnaLysis Toolkit (WebGestalt[66]) software with the following settings: *Method of Interest* = 'Overrepresentation Enrichment Analysis' (ORA) and *Functional Database* class = geneontology (GO) named 'Biological_Process_noRedundant'. Finally, we selected the *Gene ID type* 'ensemble_gene_id', to upload our orthologous data using ENSG IDs and selected genome for *Reference Set for Enrichment Analysis*. We can add 'Advanced parameters', so we keep default parameters (*Minimum Number of Genes for a Category*: 5; *Maximum Number of Genes for a Category*: 2000 and *Multiple Test Adjustment*: BH) and use the *Significant Level*: FDR 0.05.

All data are available through the dedicated website:
http://dogs.genouest.org/lncRNA.html.

References

1. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–27 (2011).
2. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–89 (2012).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
4. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–6 (2005).
5. Schein, A., Zucchelli, S., Kauppinen, S., Gustincich, S. & Carninci, P. Identification of antisense long noncoding RNAs that function as SINEUPs in human cells. *Sci. Rep.* **6**, 33605 (2016).
6. Karlsson, E. K. & Lindblad-Toh, K. Leader of the pack: gene mapping in dogs and other model organisms. *Nat. Rev. Genet.* **9**, 713–25 (2008).
7. 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).
8. Hoeppner, M. P. *et al.* An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9**, e91172 (2014).
9. 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).
10. Plassais, J. *et al.* A Point Mutation in a lincRNA Upstream of GDNF Is Associated to a Canine Insensitivity to Pain: A Spontaneous Model for Human Sensory Neuropathies. *PLoS Genet.* **12**, e1006482 (2016).

11. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* **17**, 601–14 (2016).
12. Signal, B., Gloss, B. S. & Dinger, M. E. Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs. *Trends Genet.* **32**, 620–637 (2016).
13. Johnson, R. & Guigó, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20**, 959–76 (2014).
14. Kapusta, A. *et al*. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).
15. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).
16. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
17. Wang, W. & Kirkness, E. F. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res.* **15**, 1798–808 (2005).
18. Minnick, M. F., Stillwell, L. C., Heineman, J. M. & Stiegler, G. L. A highly repetitive DNA sequence possibly unique to canids. *Gene* **110**, 235–8 (1992).
19. Lin, L. *et al*. The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* **98**, 365–76 (1999).
20. Pelé, M., Tiret, L., Kessler, J.-L., Blot, S. & Panthier, J.-J. SINE exonic insertion in the PTPLA gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Hum. Mol. Genet.* **14**, 1417–27 (2005).
21. Hédan, B. *et al*. Coat colour in dogs: Identification of the Merle locus in the Australian shepherd breed. *BMC Vet. Res.* **2**, 1–10 (2006).
22. Parker, H. G. *et al*. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325**, 995–8 (2009).
23. Guttman, M. *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–7 (2009).
24. Djebali, S. *et al*. Bioinformatics Pipeline for Transcriptome Sequencing Analysis. *Methods Mol. Biol.* **1468**, 201–19 (2017).
25. Harrow, J. *et al*. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–74 (2012).
26. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
27. Soumillon, M. *et al*. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–90 (2013).
28. Yanai, I. *et al*. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–9 (2005).
29. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).
30. Kornienko, A. E. *et al*. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol.* **17**, 14 (2016).
31. Chalmel, F. & Rolland, A. D. Linking transcriptomics and proteomics in spermatogenesis. *Reproduction* **150**, R149–57 (2015).
32. Liu, S. *et al*. Annotation and cluster analysis of spatiotemporal- and sex-related lncRNA expression in rhesus macaque brain. *Genome Res.* **27**, 1608–1620 (2017).
33. Mehravar, M. *et al*. Introduction of novel splice variants for CASC18 gene and its relation to the neural differentiation. *Gene* **603**, 27–33 (2017).
34. Paterson, C. *et al*. Temporal, Diagnostic, and Tissue-Specific Regulation of NRG3 Isoform Expression in Human Brain Development and Affective Disorders. *Am. J. Psychiatry* **174**, 256–265 (2017).
35. Andrée, B. *et al*. Isolation and characterization of the novel popeye gene family expressed in skeletal muscle and heart. *Dev. Biol.* **223**, 371–82 (2000).
36. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **Chapter 4**, Unit4.10 (2009).
37. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
38. Franke, V. *et al*. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Res.* **27**, 1384–1394 (2017).
39. Herrero, J. *et al*. Ensembl comparative genomics resources. *Database (Oxford).* **2016**, bav096 (2016).
40. Chen, Z.-H. *et al*. The lncRNA HOTAIRM1 regulates the degradation of PML-RARA oncoprotein and myeloid cell differentiation by enhancing the autophagy pathway. *Cell Death Differ.* **24**, 212–224 (2017).
41. Tripathi, V. *et al*. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **39**, 925–38 (2010).
42. Nakagawa, S., Naganuma, T., Shioi, G. & Hirose, T. Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *J. Cell Biol.* **193**, 31–9 (2011).
43. Kretschmer, A. & Tilki, D. Biomarkers in prostate cancer - Current clinical utility and future perspectives. *Crit. Rev. Oncol. Hematol.* **120**, 180–193 (2017).
44. Ke, D. *et al*. The combination of circulating long noncoding RNAs AK001058, INHBA-AS1, MIR4435-2HG, and CEBPA-AS1 fragments in plasma serve as diagnostic markers for gastric cancer. *Oncotarget* **8**, 21516–21525 (2017).
45. Hagan, J. P., O'Neill, B. L., Stewart, C. L., Kozlov, S. V. & Croce, C. M. At least ten genes define the imprinted Dlk1-Dio3 cluster on mouse chromosome 12qF1. *PLoS One* **4**, e4352 (2009).
46. Davydov, E. V. *et al*. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
47. Schüler, A., Ghanbarian, A. T. & Hurst, L. D. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* **31**, 3164–83 (2014).
48. Breschi, A., Gingeras, T. R. & Guigó, R. Comparative transcriptomics in human and mouse. *Nat. Rev. Genet.* **18**, 425–440 (2017).
49. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
50. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–5 (2013).
51. Mallory, A. C. & Shkumatava, A. LncRNAs in vertebrates: advances and challenges. *Biochimie* **117**, 3–14 (2015).
52. Albrecht, A.-S. & Ørom, U. A. Bidirectional expression of long ncRNA/protein-coding gene pairs in cancer. *Brief. Funct. Genomics* **15**, 167–73 (2016).
53. Cai, Q. & Chan, H. P. A Double Application of the Benjamini-Hochberg Procedure for Testing Batched Hypotheses. *Methodol Comput Appl Probab* **19**, 429–443 (2016).
54. Villar, D. *et al*. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**, 554–566 (2015).
55. Sun, J., Lin, Y. & Wu, J. Long non-coding RNA expression profiling of mouse testis during postnatal development. *PLoS One* **8**, e75750 (2013).
56. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **42**, 8845–60 (2014).

57. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–40 (2013).
58. Breschi, A. *et al.* Gene-specific patterns of expression variation across organs and species. *Genome Biol.* **17**, 151 (2016).
59. Qureshi, I. A. & Mehler, M. F. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nat. Rev. Neurosci.* **13**, 528–41 (2012).
60. Zhou, J. *et al.* Conserved expression of ultra-conserved noncoding RNA in mammalian nervous system. *Biochim. Biophys. Acta* **1860**, 1159–1168 (2017).
61. Carpenter, S. *et al.* A long noncoding RNA mediates both activation and repression of immune response genes. *Science* **341**, 789–92 (2013).
62. Sun, L. *et al.* Long noncoding RNAs regulate adipogenesis. *Proc. Natl. Acad. Sci. USA* **110**, 3387–92 (2013).
63. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355** (2017).
64. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, (15–21 (2013).
65. Casper, J. *et al.* The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* **46**, D762–D769 (2018).
66. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* **45**, W130–W137 (2017).

Acknowledgements

The biological samples were obtained from the 'Cani-DNA_CRB', which is part of the CRB-Anim infrastructure (http://dog-genetics.genouest.org). We would like to thank the following funding agencies: the 'Cani-DNA_CRB' [ANR-11-INBS-0003], the CNRS (Centre National de la Recherche Scientifique), the Université Rennes1 and the Brittany Region (France) (PhD funding for CLB). We thank the LUPA consortium (http://eurolupa.org) and the BROAD Institute for the sequencing. We thank Sarah Djebali of the INRA GenPhyse of Toulouse and members of the canine genetics team at IGDR-Rennes. We thank Ignacio Gonzalez for help with the analysis using mixOmics. We also thank the Genouest Bioinformatic platform, Rennes, France (https://www.genouest.org/) for providing the necessary computing resources.

Author Contributions

C.L.B., T.D. and C.H. conceived the study and wrote the manuscript. C.L.B., V.W., L.L., E.C., B.H., N.B., C.D.B., A.S.G., C.A., T.D. and C.H. provided the samples and tissues for phenotyping. Data analysis was performed by C.L.B., V.W., T.D. and C.H. All authors read, improved and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-31770-2.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

Nucleic Acids Research Advance Access published January 3, 2017

Nucleic Acids Research, 2016 **1–12**
doi: 10.1093/nar/gkw1306

FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome

Valentin Wucher[1], Fabrice Legeai[2,3], Benoît Hédan[1], Guillaume Rizk[3], Lætitia Lagoutte[1], Tosso Leeb[4], Vidhya Jagannathan[4], Edouard Cadieu[1], Audrey David[2], Hannes Lohi[5,6], Susanna Cirera[7], Merete Fredholm[7], Nadine Botherel[1], Peter A.J. Leegwater[8], Céline Le Béguec[1], Hille Fieten[8], Jeremy Johnson[9], Jessica Alföldi[9], Catherine André[1], Kerstin Lindblad-Toh[9,10], Christophe Hitte[1] and Thomas Derrien[1,*]

[1]Institut Génétique et Développement de Rennes, CNRS, UMR6290, University Rennes1, Rennes, Cedex 35043, France, [2]IGEPP, BIPAA, INRA, Campus Beaulieu, Le Rheu 35653, France, [3]Institut National de Recherche en Informatique et en Automatique, Institut de Recherche en Informatique et Systèmes Aléatoires, Genscale, Campus Beaulieu, Rennes 35042, France, [4]Institute of Genetics, Vetsuisse Faculty, University of Bern, Bern 3001, Switzerland, [5]Department of Veterinary Biosciences and Research Programs Unit, Molecular Neurology, University of Helsinki, PO Box 63, Helsinki 00014, Finland, [6]The Folkhälsan Institute of Genetics, Helsinki 00014, Finland, [7]Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 1870, Denmark, [8]Department of Clinical Sciences of Companion Animals, Faculty of Veterinary Medicine, Utrecht University, Utrecht 3584CM, the Netherlands, [9]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA and [10]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala 751 23, Sweden

Received September 23, 2016; Revised December 13, 2016; Editorial Decision December 14, 2016; Accepted: December 14, 2016

ABSTRACT

Whole transcriptome sequencing (RNA-seq) has become a standard for cataloguing and monitoring RNA populations. One of the main bottlenecks, however, is to correctly identify the different classes of RNAs among the plethora of reconstructed transcripts, particularly those that will be translated (mRNAs) from the class of long non-coding RNAs (lncRNAs). Here, we present FEELnc (FlExible Extraction of LncRNAs), an alignment-free program that accurately annotates lncRNAs based on a Random Forest model trained with general features such as multi *k*-mer frequencies and relaxed open reading frames. Benchmarking versus five state-of-the-art tools shows that FEELnc achieves similar or better classification performance on GENCODE and NONCODE data sets. The program also provides specific modules that enable the user to fine-tune classification accuracy, to formalize the annotation of lncRNA classes and to identify lncRNAs even in the absence of a training set of non-coding RNAs. We used FEELnc on a real data set comprising 20 canine RNA-seq samples produced by the European LUPA consortium to substantially expand the canine genome annotation to include 10 374 novel lncRNAs and 58 640 mRNA transcripts. FEELnc moves beyond conventional coding potential classifiers by providing a standardized and complete solution for annotating lncRNAs and is freely available at https://github.com/tderrien/FEELnc.

INTRODUCTION

The development of high-throughput RNA sequencing (RNA-seq) has revealed the presence of many RNAs in different organisms such as mammals (1–3), insects (4,5) and plants (6,7). Particularly, whole transcriptome sequencing sheds light on the pervasive transcription of the genomes with messenger RNAs (mRNAs) only representing a small fraction of the genome, outnumbered by a vast repertoire of small (miRNAs, snRNAs...) and long non-coding RNAs (lncRNAs) (1). LncRNAs, basically defined as transcripts longer than 200 nucleotides and without any protein-coding capabilities, have been involved in many aspects of normal and pathological cells. From the pioneer discovery of the *Xist* lncRNA involved in X chromosome inactivation in placental females (8) to the more recent links between lncRNAs and cancers (9,10), lncRNAs have emerged as key actors of the cell machinery with diverse modes of action such

*To whom correspondence should be addressed. Tel: +33 223236534; Fax: +33 223234478; Email: tderrien@univ-rennes1.fr

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

as gene expression regulation, control of translation or imprinting.

Following RNA sequencing, the computational reconstruction of transcripts models either by genome-guided (11,12) or *de novo* assembly (13) usually produces tens of thousands of known and novel transcript models. Among this wealth of assembled transcripts, it remains crucial to annotate the different classes of RNAs and especially to distinguish protein-coding from non-coding RNAs. To this aim, several bioinformatic tools have been developed in order to compute a coding potential score (hereafter termed CPS) used to discriminate the coding status of the RNA gene models. Broadly, they can be divided into programs using sequence alignments, either between species (14) or alignments to protein databases (15), and alignment-free software (16–18). The alignment-dependent methods, although very specific in terms of performance, are often very time- and resource-consuming. For instance, the PhyloCSF program requires a multispecies sequence alignment to predict the likelihood of a sequence to be a conserved protein-coding transcript based on the evolution of the codon substitution frequencies (14). It is thus dependent on the quality of the input alignments and may also be biased toward misclassifying species-specific or lowly conserved coding and non-coding transcripts (19). In contrast, the alignment-free methods compute a CPS only depending on intrinsic features of the input RNA sequences. One of the main features is given by the length of the longest open reading frame (ORF) (20,21) since a transcript harboring a long ORF will most likely be translated into a protein. However, the definition of the longest ORF can vary between programs, especially when it involves the strict inclusion of either or both start and stop codons. This is particularly important to model since some transcripts from reference annotations and/or newly assembled transcripts are not full-length. For instance, the number of protein-coding transcripts in the human EnsEMBL (v83) annotation (22) lacking a start codon or a stop codon is 7677 (∼10%) and 16 649 (∼25%), respectively. A complementary feature to discriminate mRNAs from non-coding RNAs is the relative frequency of oligonucleotides or k-mer (where k denotes the size of the oligonucleotide). Some tools already use k-mer frequencies but are often limited to one and/or small k-mers (generally $k \leq 6$), whereas longer k-mers could help resolve ambiguities by taking into account lncRNA-specific repeats or spatial information (23,24). Finally, common to all methods is the lack of an explicit modeling and cut-off definition for 'non-model' organisms (25), for which it can be crucial to train the programs with species-specific data and to automatically derive a CPS cut-off which provides better discriminative power.

Here, we present FEELnc, for FlExible Extraction of LncRNAs, a new tool to annotate lncRNAs from RNA-seq assembled transcripts. FEELnc is an all-in-one solution from the filtering of non-lncRNA-like transcript models, to the computation of a coding potential score and the formalization of the definition of the lncRNA classes. Based on a relaxed definition of ORFs and a very fast analysis of small and large k-mer frequencies (from k = 1 to 12), the program implements an alignment-free strategy using Random Forests (26) to classify lncRNAs and mRNAs. We benchmarked FEELnc and five existing programs (PhyloCSF (14), CPC (15), CPAT (16), PLEK (17) and CNCI (18)) using known sets of lncRNAs annotated in multiple organisms (GENCODE for human and mouse (27) and NONCODE for other species (28)), and showed that FEELnc performance metrics outperformed or are similar to state-of-the-art programs. We developed FEELnc to be used on 'non-model' organisms for which no set of lncRNAs is available by deriving species-specific lncRNA models from mRNA sequences and automatically computing the CPS cut-off that maximizes classification performances. FEELnc also allows users to provide their own specificity thresholds in order to annotate high-confidence sets of lncRNAs and mRNAs and to define a class of transcripts with ambiguous status. Finally, as part of the LUPA consortium (29), we produced 20 RNA sequencing data sets from 16 different canine tissues and applied FEELnc on the reconstructed models to annotate 10 374 lncRNA and 58 640 mRNA new transcripts from known and novel loci. We also classified lncRNAs into 5033 long intergenic non-coding RNAs (lincRNAs) and 5341 genic sense or antisense lncRNAs based on the FEELnc classifier module. The number of lncRNA transcripts detected by our data considerably expands the canine genome annotation providing an extended resource which will help deciphering genotype to phenotype relationships (30).

MATERIALS AND METHODS

Data set

For the sake of reproducibility, all data sets and scripts used to generate the benchmark files are available in Supplementary Data.

Human long non-coding and protein-coding genes were obtained from the manually curated GENCODE version 24 annotation (EnsEMBL v83 corresponding to the GRCh38 human genome assembly) selecting the long non-coding gene biotypes '*lincRNA*' and '*antisense*', and '*protein_coding*' for coding genes. From each of this set, 10 000 transcripts were extracted and further divided into two sets of 5000 transcripts, that are used for the learning and the testing steps, denoted HL and HT data sets, respectively. Importantly, only one transcript per locus was extracted for all biotypes in order not to create a bias by introducing two isoforms of the same gene in both the HL and HT sets. For mouse, we used the GENCODE version M4 annotation (EnsEMBL v79) and derived the learning and testing sets in the same way as for human (denoted ML and MT). Due to the lower number of GENCODE lncRNAs annotated in mouse compared to human, each file contains ∼2000 lncRNAs and 5000 mRNAs. For 'non-model organisms', lncRNAs belonging to the lincRNA and antisense classes (NONCODE codes 0001 and 1000, respectively) were downloaded from the latest version of the NONCODE database (NONCODE 2016) (28) while mRNAs were retrieved from the EnsEMBL database (v84). A summary of the number of mRNAs/lncRNAs per species is available in Supplementary Table S1.

Whole transcriptome sequencing of dog RNA samples (n = 20) was performed by the LUPA consortium. These biological samples, corresponding to 16 unique tissues

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

and 7 breeds, were obtained from the 'Cani-DNA CRB' biobank at the University Rennes1, CNRS-IGDR, France (http://dog-genetics.genouest.org), the biobank at University of Copenhagen, Denmark, the biobank at University of Helsinky, Finland and the Vetsuisse Biobank at University of Bern, Switzerland. The dog owners consented to the use of the data for research purposes anonymously. RNAs were extracted from tissues using the NucleoSpin RNA kit (Macherey–Nagel) according to the manufacturer's instructions. Polyadenylated RNAs were captured by oligo-dT beads and RNA-seq libraries were constructed via the Illumina TruSeq™ Stranded mRNA Sample Preparation Kit. Sequencing was done in paired-end and stranded fashion on the HiSeq2000 platform using v3 chemistry (TruSeq PE Cluster Kit v3-cBot-HS, TruSeq SBS Kit v3-HS, TruSeq SR Cluster Kit v3-cBot-HS) to a depth of about 50 million reads per tissue (Supplementary Table S2). The RNA-seq data are available in the short read archive (SRA) under NCBI bioproject PRJNA327075 and SRA accession SRP077559.

FEELnc filter module ($\text{FEELnc}_{\text{filter}}$)

The first FEELnc module aims at identifying non-lncRNA transcripts from the reconstructed transcript models given by genome-guided transcriptome assemblers such as Cufflinks (11) or more recently StringTie (12). To achieve this goal, FEELnc flags every assembled transcript that overlaps any exon of the reference annotation in sense. To deal with the plethora of input models inherent to high-depth RNA-seq experiments, the comparison of transcript intervals is parallelized through the Parallel:ForkManager Perl module. Importantly, FEELnc allows the user to parameter the percentage of overlap and also the transcript biotype (e.g. '*protein_coding*' or '*pseudogene*') to be considered from the reference annotation. Indeed, transcripts matching protein-coding exons should be flagged as they likely indicate novel mRNA isoforms. $\text{FEELnc}_{\text{filter}}$ also filters out short transcripts (default 200 nt) and can deal with single-exon transcripts depending on whether the protocol used to construct libraries is stranded or not. For instance, the module allows the removal of intergenic single-exon models as they may correspond to mapping artifacts due to repeat sequences and for which the checking of the consensus splice sites could not be assessed (27).

FEELnc coding potential module ($\text{FEELnc}_{\text{codpot}}$)

FEELnc_{codpot} predictors. The second FEELnc module aims at computing a coding potential score given the assembled sequences following transcriptome reconstruction. To deal with the incompleteness of ORF annotation where both the reference and the reconstructed transcripts may not be full-length, FEELnc computes all ORFs and annotates five ORF types from the stricter 'type 0' which corresponds to the longest ORF having both a start and a stop codon, to the more relaxed 'type 4' that is the whole input RNA sequence (see Supplementary Data for detailed description of the ORF types). Because the size of the protein-coding ORF is generally correlated with the length of the input RNA sequence, we used the *ORF coverage*, i.e. the proportion of the transcript size covered by an ORF, as the first predictor to discriminate mRNAs/lncRNAs in the FEELnc model.

The second predictor of FEELnc relies on the computation of the *multi k-mer frequencies* between mRNAs and lncRNAs. Biases in nucleotide frequencies and codon usage have already been described in the literature as important discriminative features between coding and non-coding RNAs (21,31). Within the framework of FEELnc, we developed an extremely fast and exact *k*-mer counter called KIS (for *K*-mer in short) that relies on the open-source GATB library (32). For example, KIS can compute all 6-mers (hexamers) and 12-mers of the human GRCh38 genome assembly (~3 billions *k*-mers) in ~2 min and 2 min 50 s, respectively (on a linux RedHat station with one core Intel(R) Xeon(R) CPU X5550 @ 2.67GHz). Due to the high speed of KIS, a major contribution of this work was to be able to combine different lists of *k*-mers, including longer *k*-mers, in order to better discriminate lncRNAs from mRNAs. Specifically, we assigned a score for each sequence *K* (e.g. TGC) of size *k* (e.g. 3) based on the occurrence of this sequence in each predicted mRNA ORF sequence and lncRNA whole sequence from the learning data set. This score S_K^k, similar to the one used in Claverie *et al.* in (33), is computed for each *K* of size *k* as follows:

$S_K^k = \frac{F_K^m}{F_K^m + F_K^{lnc}}$, with F_K^m and F_K^{lnc} the observed frequencies of *K* in mRNA ORFs and in lncRNA sequences for the two learning sets, respectively. Note that these *k*-mer profiles are made on a subset of the learning set (10% by default) and the transcripts used to compute the *k*-mer profiles are removed from the random forest model in order to avoid overfitting.

Once the *k*-mer profiles are made, i.e. all S_K^k have been computed for all *k*-mer, FEELnc associates a *k*-mer score V_X^k for each remaining ORFs *X* as follows:

$V_X^k = \frac{\sum_{K=1}^{4^k} S_K^k \times N_K^X}{\sum_{j=1}^{4^k} N_j^X}$, with N_K^X the number of occurrences of *K* in the ORF *X* and $\sum_{j=1}^{4^k} N_j^X$ the total number of *k*-mer of size *k*. Using this scoring method, a *k*-mer score is associated to each sequence for each *k*-mer size selected in the model.

FEELnc coding potential also uses the total *RNA sequence length* as a predictor of the model since lncRNAs have been shown to be significantly shorter than mRNAs (34,35). For illustration purposes, a distribution of the FEELnc predictor scores with 'type 3' ORF (i.e. longest ORF having either or both a start and a stop codon) and multi *k*-mer scores with *k* in {1, 2, 3, 6, 9, 12} is given in Supplementary Figure S1 for the 5000 mRNAs and 5000 lncRNAs of the HL data set.

Random forest classification and optimized coding potential cut-offs. The aforementioned predictor scores are incorporated into a machine learning method—Random Forest (RF) (26)—that computes a coding potential score (CPS) for each input training transcripts. As also shown by others with respect to lncRNAs annotation, RF often outperforms other machine learning techniques especially due to the random sampling of features to build the ensemble of trees (36,37). In addition, our RF model which is based on the randomForest R package (38), can deal with im-

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de lIRMAR on January 10, 2017

balanced training set by down-sampling the majority class (most likely mRNAs in many organisms). In fact, the CPS in our RF model corresponds to the proportion of all trees (500 trees by default) which 'vote' for the input sequence to be coding or non-coding. A proportion close to 0 will indicate a non-coding RNA and close to 1 an mRNA.

To define an optimal CPS cut-off, FEELnc automatically extracts CPS that maximizes both sensitivity (Sn) and specificity (Sp) (see performance section) based on a 10-fold cross-validation. Using the ROCR package (39), FEELnc provides users with a two-graph ROC curve in order to display the performances of the model and to visualize the optimal CPS (Figure 1 A).

Even if this approach aims at providing the highest performances, it could sometimes misclassify coding and/or non-coding transcripts whose CPS is closed to the optimized threshold (40). To take this into account, FEELnc allows fixing two minimal specificity cut-offs for lncRNAs and mRNAs (this approach is termed 'two cut-offs'). This naturally leads to the annotation of two high-confident classes of lncRNAs and mRNAs and also to the definition of a third class of ambiguous transcripts (i.e. transcripts whose the CPS is between the two cut-offs) that are named TUCp (34,40) for Transcripts of Unknown Coding potential (Figure 1A).

FEELnc without long non-coding training set. One issue when using machine learning algorithms is the requirement of both a positive and negative sets (here mRNA and lncRNA) to train the model. While the former is often available for most organisms, the latter is usually not especially for non-model organisms (25). To model non-coding RNAs in the absence of a true set of lncRNAs, we assessed three strategies called *intergenic*, *shuffle* and *cross-species*. As DNA composition varies between species, a first naive approach consists in extracting decoy sequences from the genome of interest to model species-specific non-coding sequences. More precisely, we extracted random intergenic sequences of length L (L is given by the distribution of the mRNA sizes) as the non-coding training set. The shuffle strategy employs a more sophisticated method which is based on the idea that lncRNAs are derived from 'debris' of protein-coding genes (41,42), as exemplified by the *Xist* lncRNA that emerged from the disruption of the mRNA gene *Lnx3* (43). To this end, we shuffled mRNA sequences from the reference annotation using the Ushuffle program (44), while preserving a given k-mer frequency of the input sequences. Note, however, that shuffling sequences and preserving frequencies for one fixed k-mer does not constrain the frequencies of the other k-mers. In addition, when k increases, it is possible that Ushuffle could not permute some input sequences because of the constraint to preserve the given k-mer frequencies. Finally, the cross-species strategy makes use of lncRNA sets annotated in other species to extract non-coding predictors and train the RF model. For the latter strategy, we trained the FEELnc model using human mRNAs and species-specific lncRNAs (NONCODE lncRNA catalogues being available in 13 different species). For all strategies, we assessed the performance on the HT data sets.

Performance evaluation. We evaluated the performance of FEELnc and five other state-of-the-art programs: CNCI (version 2 Feb 28, 2014) (18), CPC (version 0.9-r2) (15), CPAT (version 1.2.1) (16), PhyloCSF (version 20121028-exe) (14) and PLEK (version 1.2) (17), by computing classical performance metrics:

- Sensitivity (Sn) or True Positive Rate $= \frac{TP}{TP+FN}$;
- Specificity (Sp) or True Negative Rate $= \frac{TN}{FP+TN}$;
- Precision (Prec) or Positive Predicted Value $= \frac{TP}{TP+FP}$;
- Accuracy (Acc) $= \frac{TP+TN}{TP+FP+TN+FN}$;

With TP: True Positive, TN: True Negative, FP: False Positive and FN: False Negative.

In addition, we used two complementary metrics in order to capture the global performance of the tools in a single measure (45):

- F-score $= 2 \times \frac{precision \times sensitivity}{precision+sensitivity}$, which is a statistic measuring the harmonic mean of precision and sensitivity;
- MCC (Matthews Correlation Coefficient) $= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, which is particularly useful when the two classes are of very different sizes (which is often the case for mRNAs and lncRNAs in non-model organisms) and which could be seen as a correlation coefficient between the true classes and the predicted classes (46).

For each performance metric, we considered lncRNAs as the negative class and mRNAs as the positive class. Note that, as a binary classification, the mRNA specificity corresponds to the lncRNA sensitivity (and conversely). Moreover, for the CPAT and PLEK programs, which allow training their models, we used species-specific set of mRNAs/lncRNAs for training (called $CPAT_{train}$ and $PLEK_{train}$). Nevertheless, contrary to FEELnc, PLEK and CPAT required us to a priori extract the CDS of the mRNA input file to learn the coding parameters. For $CPAT_{train}$, we referred to the optimal CPS cut-off mentioned on their website (16) to discriminate between coding and non-coding RNAs. A detailed description of the command lines used to run each program is given in Supplementary Data.

FEELnc classifier module ($FEELnc_{classifier}$)

Given a known reference annotation, it is essential to classify newly annotated lncRNAs based on their closest annotated transcripts. It will potentially guide researchers towards functional annotation and relationships between lncRNAs and their annotated partners (lncRNA/mRNA pairs for instance). For this purpose, the FEELnc classifier module ($FEELnc_{classifier}$) employs a sliding window strategy (whose length is fixed by the user) around each lncRNAs to report all the reference transcripts located within the window. Not only does the FEELnc classifier annotate lincRNAs and antisense lncRNAs but it also formalizes the definition of lncRNA subclasses with respect to annotated transcripts (Figure 1B). First, these rules involve the *direction* (sense or antisense) and the *type* of interactions (genic or intergenic). Then, within each *type* of interaction, a *subtype* level allows to narrow down the classification (e.g. divergent for lincRNAs or containing for genic lncRNAs). Finally, a *location* level is added informing about the position of the

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

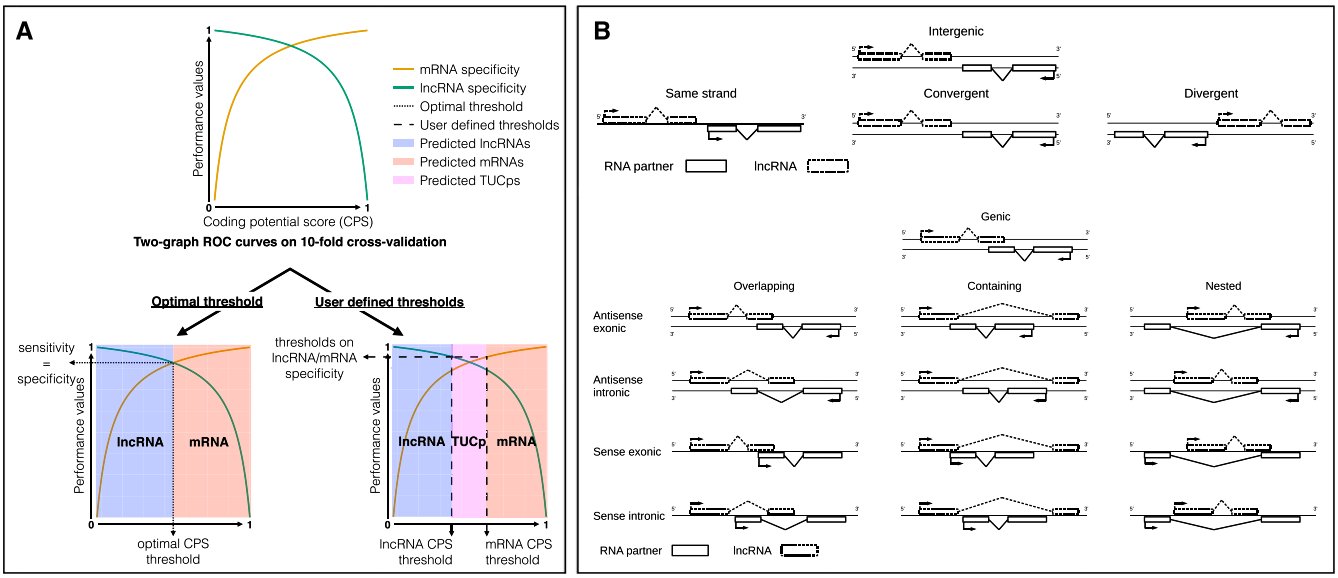

Figure 1. FEELnc$_{codpot}$ and FEELnc$_{classifier}$ description. (**A**) Two graph ROC curves for automatic detection of optimized CPS threshold and user specificity threshold, the latter defining two conservative sets of lncRNAs and mRNAs and a class of transcripts with ambiguous biotypes termed TUCp (Transcripts of Unknown Coding potential). **B**) Sub classification of intergenic and genic lncRNA/transcripts interactions by the FEELnc$_{classifier}$ module.

lncRNA with respect to the annotated transcripts (e.g. upstream for lincRNA or exonic for genic lncRNAs). In addition, the FEELnc$_{classifier}$ can be used with all transcript biotypes (e.g. short ncRNAs such as snoRNAs) from a reference annotation and therefore is capable of annotating lncRNAs that are host genes for short RNAs (47) (hence the '*genic sense exonic*' class).

Because one lncRNA could belong to different classes depending on which reference transcript is considered, our approach reports all interactions within the defined window and defines a best partner transcript using the following priorities: for genic lncRNAs, the exonic class has priority over the intronic class and the intronic over the containing, while for lincRNAs, the nearest reference transcript is selected.

Reads mapping and transcript model reconstruction of canine RNA-seq samples

The processing of RNA-seq reads from the mapping of the reads to the transcript model reconstruction was performed using standard bioinformatic pipeline (48). Such genome-guided transcript model reconstruction has already been validated as for instance in the canine genome (49). Briefly, the mapping of the canine RNA-seq reads was done using the STAR v2.5.0a program (50) while Cufflinks v2.2.1 (11) was used to reconstruct transcript models for each sample separately using the dog genome annotation by EnsEMBL and by the Broad (49) as a guide. Finally, the cuffmerge tool, from the Cufflinks package, was used to compute a single consensus file with all reconstructed transcript models (all command lines and parameters used for each tool are also provided in Supplementary Data).

RESULTS

FEELnc modules to annotate lncRNAs

Starting from assembled transcripts and a reference annotation, the FEELnc pipeline is composed of three independent modules to classify and annotate lncRNAs (Figure 2). The FEELnc$_{filter}$ module filters out input transcript models reconstructed via a genome-guided assembly strategy that do not correspond to potential novel lncRNA candidates. Due to its ability to take into account reference transcript biotypes (See Methods), the module allows to keep novel transcript models overlapping other referenced ncRNAs for instance, as these models may correspond to long non-coding RNAs that are host genes for small RNAs (51). After the FEELnc$_{filter}$ module, the remaining transcripts are thus candidates to be new lncRNAs or mRNAs.

The second module (FEELnc$_{codpot}$) computes a coding potential score for every candidate transcript based on a RF model trained with several predictors such as ORF coverage, multi *k*-mer frequencies and RNA sizes (see Methods for details about predictors). Using the gold-standard GENCODE human learning set (HL) for training, we evaluated the performance of FEELnc on 5000 lncRNAs and 5000 mRNAs of the test data set with respect to the five ORF types and multi *k*-mer combinations. Remarkably, we observed that 'type 1' and 'type 3' ORFs, which extract the longest ORF even in the absence of stop codon, consistently display better achievements (mean MCC = 0.816) than 'type 0' and 'type 2' ORFs (mean MCC = 0.67 and 0.68, respectively) whichever combination of *k*-mers is considered (Supplementary Figure S2). In addition, the multi *k*-mer strategy improves the performance of the program with a MCC performance starting at 0.80 when only using *6*-mers but reaching 0.85 with a combination of {*1, 2, 3, 6, 9,*

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

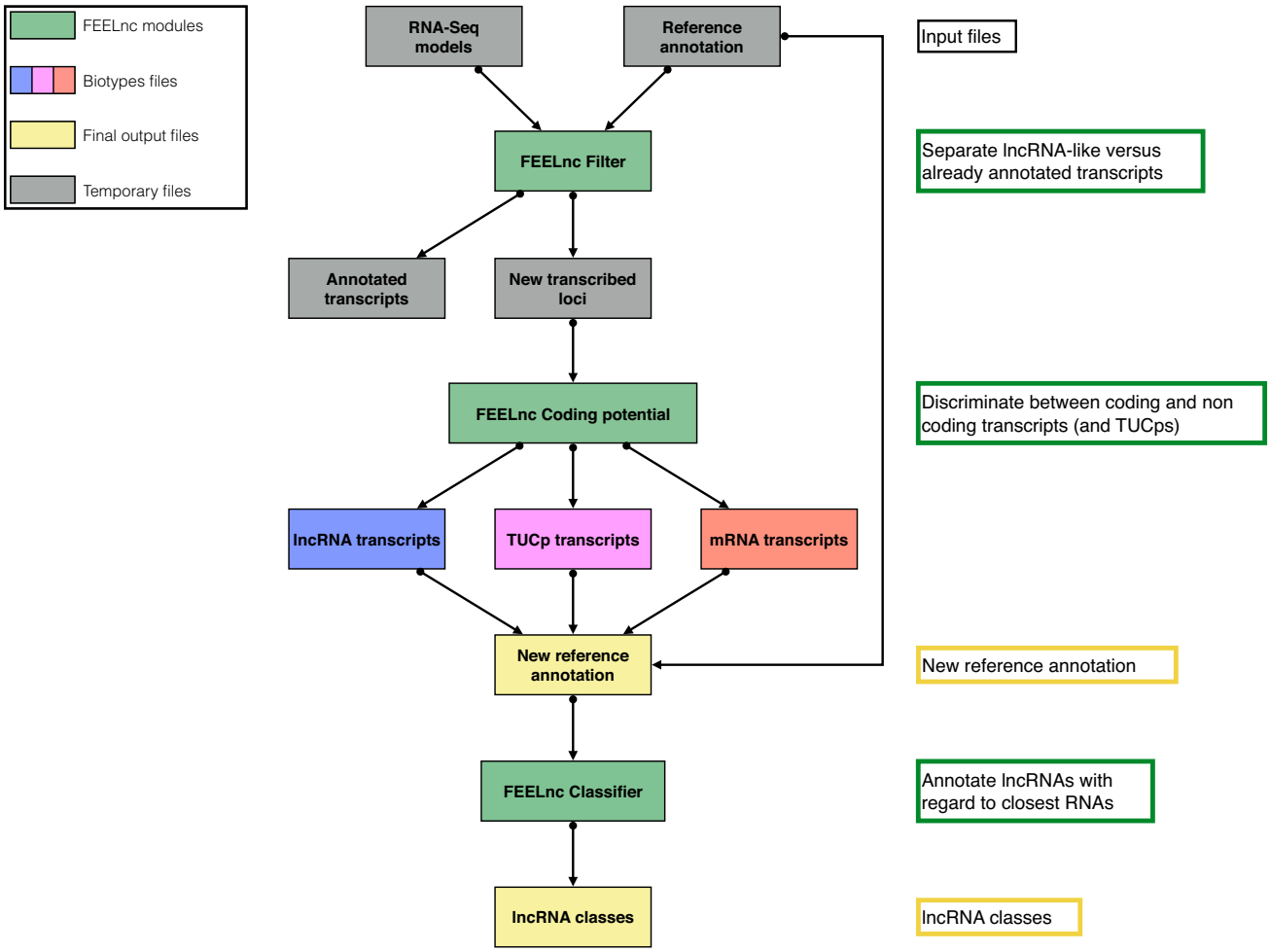

Figure 2. General overview of the FEELnc pipeline. The FEELnc filter module ($FEELnc_{filter}$) identifies newly assembled RNA-seq transcripts and removes non-lncRNA transcripts. The FEELnc coding potential module ($FEELnc_{codpot}$) computes a coding potential score (CPS) and automatically defines the optimal CPS score cut-off to discriminate lncRNAs versus mRNAs (and eventually TUCPs). The FEELnc classifier module ($FEELnc_{classifier}$) annotates lncRNA classes based on RNA partners from the reference annotation.

12}-mers (with a fixed ORF type 3, Supplementary Figure S2).

In addition to measuring performance, we conducted several evaluations to assess the robustness of the FEELnc predictions. First, we showed that FEELnc was not biased by unbalanced or low numbers of transcripts in the input training set with sensitivity and specificity values higher than 0.9 using only 400 mRNAs and lncRNAs (Supplementary Figure S3). Second, FEELnc performed similarly or better than other methods to classify very small or long mRNAs and lncRNAs (Supplementary Figure S4). Third, in order to model incomplete RNA reconstructions, we removed the 10%, 25% and 50% of either 5′-end or 3′-end of transcript sequences from the HT data set. Even if $FEELnc_{codpot}$ performance decreased with increasing degradation percentages (MCCs = 0.827, 0.749 and 0.53 for 10%, 25% and 50%, respectively), it performed better than other tested methods (Supplementary Figure S5). Fourth, we observed similar high performance with mouse GENCODE data set ML and MT (see materials) composed of 5000 mRNAs and 2000 lncRNAs where FEELnc achieves 0.938 in sensitivity, 0.941 in specificity and an MCC of 0.856. Finally, among the manually curated list of 35 'well-characterized' lncRNAs from Chen *et al.* (52), FEELnc correctly classifies 33 (95%) of them as non-coding. The two discordant lncRNAs are *PWRN1*, which has a CPS (0.374) just above the optimized cut-off (0.372), and *FIRRE,* which exhibits the highest CPS (0.58 when the median CPS for the 35 lncRNAs is 0.088). This high coding potential score for *FIRRE* lncRNA can be explained by a large and complete ORF (546 nt, i.e. 58% of the total RNA sequence) and a high level of sequence similarity with the *FAM195A* protein-coding gene.

The third FEELnc module ($FEELnc_{classifier}$) formalizes the annotation of lncRNAs based on neighboring genes in order to predict lncRNA functions and RNA partners (see classes in methods). To illustrate the outcome of the $FEELnc_{classifier}$, we applied it on the human EnsEMBL v83 annotation composed of 24 659 lncRNAs ('*lincRNA*' and '*antisense*' biotypes) and 79 901 mRNA transcripts ('*protein_coding*' biotype). For instance, $FEELnc_{classifier}$ anno-

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

tates 5544 lncRNAs as '*intergenic antisense upstream*' which correspond to divergent lincRNAs (i.e. transcribed in head to head orientation with the RNA partner) among which 28.3% (n = 1572) are less than 1kb from their mRNA partner Transcription Start Sites (TSSs). This class directly pinpoints to lincRNAs potentially sharing a bi-directional promoter with their mRNA partners (53,54). On the other hand, 408 lincRNAs located less than 5kb from their mRNA partner, belong to the '*sense intergenic upstream*' class and may correspond to dubious lncRNAs that are actually 5′UTR extensions of the neighboring protein-coding RNAs. $\text{FEELnc}_{\text{classifier}}$ also annotates 5006 lncRNAs in the '*antisense exonic*' class as potential candidates for complementary interactions with the mRNA transcribed in opposite direction (55,56).

Benchmarking FEELnc and existing tools

We next compared the performance of FEELnc with five state-of-the-art programs either alignment-free (CPAT, CNCI and PLEK) or alignment-based (PhyloCSF and CPC). Similarly to FEELnc, we used the balanced human learning data set (HL) composed of 5000 lncRNAs and 5000 mRNAs (see Methods) to construct the models for CPAT and PLEK (denoted $\text{CPAT}_{\text{train}}$ and $\text{PLEK}_{\text{train}}$). We also used default pre-built models for PLEK and CPAT although some of the transcripts from the human GENCODE data set test file (HT) could have been used for building these models. For all programs, performance metrics were calculated according to the HT data set. This showed that FEELnc had the highest classification power (AUC, i.e. Area Under the Curve value = 0.97) compared to the others tools as illustrated by the ROC curves (Figure 3A). Accordingly, FEELnc displays the highest sensitivity (0.923) and the second highest specificity (0.915) among all tools while PLEK displays the highest specificity (0.985) and precision (0.981). In general, alignment-based methods have lower classification metrics than alignment-free programs as they usually depend on the quality of the input cross-species alignments or the completeness of species-specific protein databases. Finally, FEELnc also shows the highest classification accuracy (0.919), F-score (0.919) and MCC values (0.838) indicating that it performs well on the human GENCODE data set in comparison to other tools (Table 1).

We further investigated the performance on the mouse data sets composed of 2000 lncRNAs and 5000 mRNAs in order to replicate the analysis in another organism using an unbalanced data set. For this benchmark, we included the same programs except PhyloCSF due to the labor-intensive task to extract input cross-species multiple alignments. Again, FEELnc displays the highest classification accuracy, F-score and MCC (Table 2) while CPC shows the best specificity (0.992) and precision (0.996) despite a weak sensitivity (0.744). As in human, the CPAT program performs well even if we consider both the re-trained and prebuilt models. Interestingly, CPAT used with the trained mouse model has only slightly better performance than used with the human prebuilt model. This suggests that within-species training achieves relatively few performance gains compared to cross-species training with a closely related model species.

Finally, we assessed the computational time of each program including the time required for computing the model for training-based tools on a linux RedHat station (Intel(R) Xeon(R) CPU X5550 @ 2.67GHz). FEELnc took ∼46 min to classify the 10 000 human lncRNAs and mRNAs while PLEK was the fastest (6 min as compared to ∼10 h when we trained its model) and CPC the longest (∼2 days; with default parameters).

Annotating lncRNAs without a species-specific training set of lncRNAs

In the absence of a species-specific lncRNAs set, machine learning strategies require to simulate non-coding RNA sequences to train the model. In order to evaluate the *intergenic* and *shuffle* strategies on the human training sets (see Materials and Methods), we computed their predictor scores in comparison with the true set of 5000 HL lncRNAs. For the *shuffle* strategy, it is essential to determine *a priori* which given *k*-mer frequencies should be preserved by Ushuffle to maximize classification accuracy. We thus shuffled HL mRNA sequences for different sizes of *k* and showed that preserving *7*-mer frequencies gave the best MCC values on the HT set while sustaining a high number of permuted sequences (Supplementary Figure S6). We then compared the *intergenic* versus *shuffle*-derived lncRNAs and observed that the cumulative distribution of FEELnc predictor values for the *shuffle* strategy tended to be closer to the one observed in the true set of human lncRNAs compared to the *intergenic* approach (Figure 3B). This result was confirmed by assessing performance of these two strategies on the HT data set where the *shuffle* method outperformed the *intergenic* approach (MCCs = 0.768 versus 0.646) as compared to true set of lncRNAs (MCC = 0.846) (Supplementary Figure S7).

As described above, FEELnc can be used in a stringent mode in order to distinguish high-confidence sets of lncRNAs and mRNAs. To this end, we also applied the 'two cut-offs' option for both strategies with increasing specificity thresholds (0.93, 0.96, 0.99 for both mRNAs and lncRNAs) as compared to the automatic optimal CPS cut-off defined previously (Supplementary Figure S7). With these cut-offs, we also observed higher performances for the *shuffle* approach (MCCs = 0.823, 0.881 and 0.943) versus *intergenic* (MCCs = 0.646, 0.654 and 0.785) (Supplementary Figure S7). Moreover, the greater interest of the *shuffle* approach could be appreciated by the lower variability between sensitivity and specificity metrics as defined within the FEELnc methodology compared to the *intergenic* approach (Supplementary Figure S7).

In order to directly evaluate FEELnc performance for 'non-model' organisms, we used the FEELnc *shuffle* strategy where the protein-coding predictors were learnt on species-specific mRNAs and the non-coding predictors on species-specific mRNAs shuffled by Ushuffle (with preserved *7*-mer frequencies). All tests were further assessed on the catalogues of lncRNAs annotated in the NONCODE database. We also compared the performance with CNCI knowing that NONCODE uses both CNCI and matching protein-coding coordinates from RefSeq database to remove all ncRNAs annotated as protein-coding. In Supple-

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

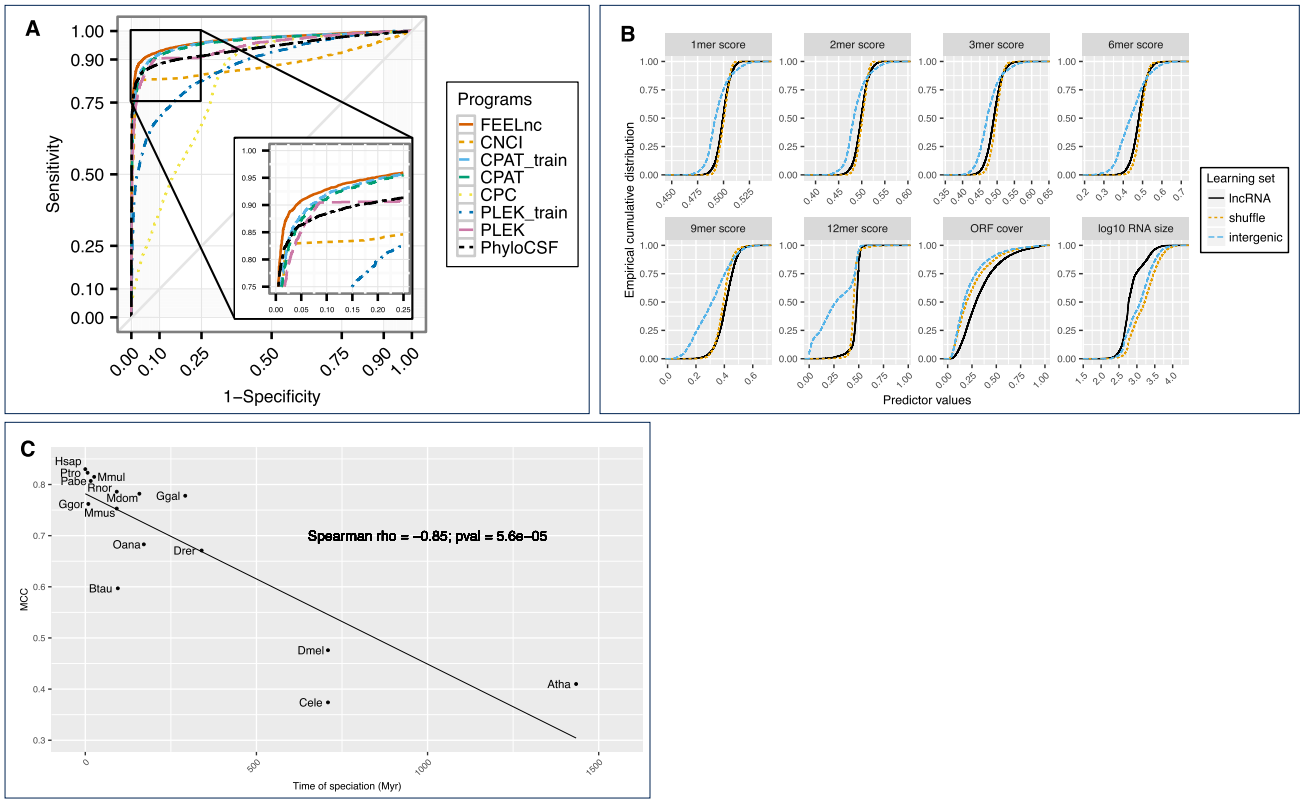

Figure 3. FEELnc performance against coding potential tools and with shuffle, intergenic and cross-species approaches. (**A**) ROC curve analysis of FEELnc versus coding potential tools based on GENCODE human data set (HT). (**B**) Empirical cumulative distribution of $\text{FEELnc}_{\text{codpot}}$ feature scores with the true set of human lncRNAs ('*lncRNA*') in comparison with the '*shuffle*' and '*intergenic*' methods. (**C**) $\text{FEELnc}_{\text{codpot}}$ MCC values tested on human HT set and trained using human mRNAs and species-specific NONCODE lncRNAs (cross-species). The x-axis represents the time of speciation between human and NONCODE species as given in (69). Species abbreviations are the following: Atha: Arabidopsis; Btau: Cow; Cele: Nematode; Dmel: Fly; Drer: Zebrafish; Ggal: Chicken; Ggor: Gorilla; Hsap: Human; Mdom: Opossum; Mmul: Rhesus; Mmus: Mouse; Oana: Platypus; Pabe: Orangutan; Ptro: Chimpanzee; Rnor: Rat.

Table 1. Tools performance on the GENCODE human data sets. Bold-underlined values correspond to the highest values of each metrics. CPAT_train and PLEK_train correspond to program versions trained with the HL data set. Programs are sorted by MCC values

HUMAN data set	Program	Sensitivity	Specificity	Precision	Accuracy	F-score	MCC
	FEELnc	**0.923**	0.915	0.916	**0.919**	**0.919**	**0.838**
	CPAT	0.899	0.924	0.922	0.912	0.910	0.823
	CPAT_train	0.920	0.901	0.903	0.910	0.911	0.821
	CNCI	0.829	0.979	0.975	0.904	0.896	0.817
	PLEK	0.732	**0.985**	**0.981**	0.858	0.838	0.741
	PhyloCSF	0.906	0.802	0.820	0.854	0.861	0.712
	PLEK_train	0.582	0.960	0.936	0.770	0.718	0.584
	CPC	0.699	0.739	0.728	0.719	0.713	0.438

Table 2. Program performances on the GENCODE mouse data sets. Bold-underlined values correspond to the highest values of each metrics. Programs are sorted by MCC values

MOUSE data set	Program	Sensitivity	Specificity	Precision	Accuracy	F-score	MCC
	FEELnc	0.938	0.941	0.976	**0.939**	**0.956**	**0.856**
	CPAT_train	**0.950**	0.880	0.952	0.930	0.951	0.828
	CPAT	0.892	0.960	0.982	0.911	0.935	0.806
	CNCI	0.857	0.972	0.987	0.890	0.918	0.772
	CPC	0.744	**0.992**	**0.996**	0.815	0.852	0.667
	PLEK	0.710	0.913	0.954	0.768	0.814	0.564
	PLEK_train	0.630	0.891	0.936	0.704	0.753	0.470

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

mentary Table S3, we showed that FEELnc without annotated lncRNAs achieved good classification metrics in many species (MCCs ranging from 0.70 for rat to 0.903 for cow) and even better than CNCI in five diverse species (nematode, fly, gorilla, orangutan and rhesus macaque).

As a third approach to model non-coding sequences, we sought to use *cross-species* lncRNAs for learning FEELnc model parameters. To this end, NONCODE lncRNAs from 13 diverse organisms were used to serve as a proxy for modeling human lncRNAs, and evaluation metrics were then performed on the HT data set. As expected, FEELnc performance is negatively correlated with respect to the evolutionary distance between human and NONCODE species (Spearman rho = −0.85; *P*-value = 5.6e-05) with MCC values of 0.374 when using *Caenorhabditis elegans* lncRNAs to 0.823 with chimpanzee lncRNAs (Figure 3C and Supplementary Table S4). This probably reflects the variability in term of lncRNA sequence conservation (and thus in non-coding *k*-mer frequencies) between human and NONCODE species (35,52). Interestingly, the *shuffle* strategy showed a MCC of 0.748, which corresponds to the performance obtained with species that diverged about 100 million years ago. This indicates that it constitutes an interesting approach when no lncRNAs from closely related species are available.

Application to identify an extended catalog of canine lncRNAs

Within the framework of the LUPA consortium (29,57), we performed 20 whole transcriptome sequencing experiments of 16 canine tissues (Supplementary Table S1). After QC, ∼1, 3 billions reads were mapped onto the CanFam3 genome assembly using the STAR mapper (50). Cufflinks (11) was further used to reconstruct the transcript models in each tissue separately guided by a consensus reference annotation given by the Broad (49) and EnsEMBL v83 (22) annotations (called CanFam3.1). Then, the cuffmerge tool (Cufflinks package) merged the tissue samples files into a single GTF file containing 211 794 transcript models of more than 50 000 gene models. In order to annotate new transcribed loci, we used the $\text{FEELnc}_{\text{filter}}$ module to flag all transcripts which overlap exons (in sense) from the reference annotation, single exonic intergenic transcripts and transcripts with a size below 200 nt. A total of 5523 remaining candidate transcripts were then analyzed by the $\text{FEELnc}_{\text{codpot}}$ module. The 'two cut-offs' option was run with a minimal specificity threshold fixed at 0.93 in both biotypes. These cut-offs allow leveraging the number of ambiguous transcripts (TUCps) while optimizing classification specificity (see Materials and Methods). This analysis identified 3822 novel lncRNA transcripts, 477 new mRNA transcripts and 884 TUCps.

Because we also aimed at developing a comprehensive annotation of novel transcript isoforms, we used $\text{FEELnc}_{\text{codpot}}$, with the exact same parameters as before, on the subset of assembled models overlapping (but not included in) the current CanFam3.1 annotation (n = 69 602) (See Supplementary Methods for details). In addition, since a novel transcript isoform could merge two or more genes from the reference annotation, we defined rules to avoid the merging of transcripts having different biotypes by removing these incompatible transcripts (see Supplementary Methods for details). This resulted in the annotation of 67 312 transcripts with more mRNA isoforms (58 163) compared to lncRNAs (6552) and TUCps (2597).

Combining the results of these analyses with the reference annotation, the new canine annotation that we called CanFam3.1-plus, includes a total of 36 237 loci (with 3145 new loci) and 189 114 transcripts (with 10 374 lncRNA and 58 640 mRNA newly annotated transcripts) (Supplementary Table S5).

This study improves the dog genome annotation in several aspects. First, we extended the number of mRNA transcripts by 50% and doubled the number of annotated lncRNA transcripts. Second, we found that novel lncRNA and mRNA transcripts were longer in terms of number of exons, CDS/UTRs and RNA sizes compared to CanFam3.1 transcripts suggesting a more complete reconstruction of their gene structures (Supplementary Table S6). Also, the number of isoforms per gene locus has been expanded to ∼2.2 and ∼7.2 for lncRNAs and mRNAs, respectively. Third, using STAR and RSEM (58) programs to quantify transcripts expression levels, we found that 86% of novel lncRNAs have a TPM (transcript per million) value higher than 0.5 in at least one of the 20 tissues highlighting a robust set of new lncRNAs. Finally, among the novel canine lncRNA genes, ∼15% are also found as non protein-coding genes in the human GENCODE annotation by using the EnsEMBL compara EPO alignments (59). The CanFam3.1-plus now annotates, for instance, three CAncer Susceptibility Candidates lincRNAs (*CASC* lincRNAs) such as the *CASC9* lincRNA located on canine chr29:23,554,585-23,605,371 and involved in esophageal squamous cell carcinoma (60). Other examples include the well-described *MALAT1* cancer-associated lincRNA (61) which was considered as an unclassified non-coding transcript in CanFam3.1 and the *IFNG-AS* antisense lncRNA involved in T-cell differentiation (62).

Finally, by employing the $\text{FEELnc}_{\text{classifier}}$ on all CanFam3.1-plus lncRNA transcripts (see Supplementary Table S4), we annotated 8209 lincRNAs. Among them, 1279 are located at a distance smaller than 5 kb and transcribed in a divergent orientation from their neighbor mRNA which could suggest potential canine bi-directional promoters (54). For genic lncRNAs, $\text{FEELnc}_{\text{classifier}}$ identified 5085 antisense exonic lncRNAs as possible candidates for sense-antisense regulation by sequence complementarity (55). The CanFam3.1-plus annotation constitutes a new resource that will help identifying lncRNA candidates for understanding genotype to phenotype relationships.

DISCUSSION

In this study, we designed a new program to identify and annotate lncRNAs called FEELnc for FlExible Extraction of Long non-coding RNAs. Using the gold-standard GENCODE annotation in human and mouse (27), we showed that FEELnc performs well to discriminate long non-coding versus protein-coding RNAs. Most probably, FEELnc includes predictors (multi *k*-mer frequencies and ORF coverage) that are general enough to cap-

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

ture all lncRNAs classes whereas alignment-based methods will be biased toward misclassifying species-specific transcripts or coding transcripts that are not referenced in peptide databases. The integration of a Random Forest-based model in FEELnc contributes to the good achievements of the tool because of the intrinsic properties of randomly subsampled features thereby encompassing diverse lncRNA characteristics. In recent years, major advances have been made in the machine learning field (63) allowing the programs to cope with thousands of parameters simultaneously. Thus, it would thus be of great value to investigate deep learning approaches to coding and non-coding RNA annotations.

The contribution of FEELnc is not only limited to provide high classification performance metrics on models organisms since the tool is also accompanied by several modules and options that enable fine-tuning and precisely adjusting lncRNA annotations for any species of interest. To our knowledge, it is the first tool that allows users to annotate conservative sets of lncRNAs and mRNAs by automatically fixing their own specificity thresholds (40). Second, FEELnc can be used for any given species even in the absence of a lncRNA training set due to the possibility to model species-specific lncRNAs. For instance, the *shuffle* strategy only requires species-specific protein-coding sequences and it is thus suitable for analyses without a reference genome assembly, which is still the case for many non-model organisms. Third, we expect the FEELnc classifier module to be of great interest to researchers in order to automatically annotate novel lncRNAs thus directly providing candidate pairs of lncRNA and mRNA partners to be investigated for experimental validations.

For non-model organisms, we have shown that FEELnc performs similarly to CNCI although the benchmark was done on NONCODE lncRNAs that were *a priori* filtered by the CNCI tool. As for human and mouse where the manually curated GENCODE annotation is considered as a standard, this stressed the importance of also defining gold-standard sets of lncRNAs/mRNAs to correctly evaluate programs for 'non-model' organisms. This is envisaged within the framework of collaborative projects such as FAANG, the Functional Annotation of Animal Genomes project (30).

Although the purpose of FEELnc is not to assemble transcripts from RNA-seq data, the program relies on the correct modeling of transcripts. For weakly expressed mRNAs, the corresponding reconstructed models might not be full-length and thus may also introduce a bias for the FEELnc sequence length predictors. However, the benchmarking on shortened sequences (Supplementary Figure S5) showed that FEELnc performs well even if 25% of either transcript-end sequences are removed. With the availability of very long reads from third generation sequencing technologies (64), the issue will not so much consist in annotating full-length transcripts but rather in taking into account the high error rate inherent to these technologies, which could lead to the misannotation of the correct ORF. Using a specific FEELnc option, preliminary tests on degraded transcript sequences showed that it raised good evaluation metrics (Sn = 0.87, Prec = 0.78) when the computation of multi *k*-mer frequencies is done on the entire transcript sequence (i.e independently of the ORF annotation) and the ORF coverage is removed from the predictors (see Supplementary Table S7).

Finally, we illustrated the usefulness of FEELnc on the dog transcriptome for which 20 RNA-seq samples were sequenced in the frame of the LUPA consortium. The biological relevance of this expanded canine resource can be illustrated by the increased transcript and CDS sizes as well as higher exon numbers. Although this improved annotation will facilitate the identification of genotype to phenotype associations, it will still need further investigations given the plethora of expressed biotypes beside lncRNAs and mRNAs. For instance, one could consider annotating transcribed (processed or unprocessed) pseudogenes or enhancer RNAs by using FEELnc as a multiclass classifier instead of a binary classifier (e.g. coding versus non-coding). Indeed, pseudogenes that recently derived from a protein-coding gene should harbor multi *k*-mer frequencies similar to their parent mRNAs but without long ORFs. Their CPS would then be gathered in an intermediate class between lncRNAs and mRNAs. For instance, FEELnc identifies that the *FIRRE* lncRNA has an intermediary CPS score of 0.58 that is supported by a relatively long ORF of 152 AA and significant percentage of sequence similarity with the *FAM195* protein-coding genes. Interestingly, it also appeared that the fourth exon of *FIRRE* is almost completely embedded in the *MCRIP2P1* pseudogene, the latter deriving from *FAM195* mRNA. This highlights the potential use of FEELnc to annotate pseudogene-derived lncRNAs (65,66) for which the CPS score should be in between lncRNAs and mRNAs. Similarly, enhancer RNAs defined as non-coding transcripts derived from enhancer elements (67,68) should harbor small ORFs and specific patterns of *k*-mers corresponding to their transcription factor binding sites, which would then be caught by FEELnc predictors.

Altogether, FEELnc provides a standardized and exhaustive protocol to identify and annotate lncRNAs.

AVAILABILITY

FEELnc is implemented in Perl and R (and KmerInShort in C++) and is available throw a github: https://github.com/tderrien/FEELnc. A UCSC track hub for the extended canine annotation CanFam3.1-plus is available using the following URL:

http://tools.genouest.org/data/tderrien/canFam3.1p/annotation/trackhub/canfam3.1p_trackhub/hub.txt with transcripts from the original CanFam3.1 annotation in blue, new isoforms from CanFam3.1 genes in green and transcripts from CanFam3.1-plus genes in red. A guideline is available in the Supplementary Data for running FEELnc analyses with or without a reference genome.

ACCESSION NUMBERS

The RNA-seq data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under NCBI bioproject PRJNA327075 and SRA accession SRP077559.

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the Cani-DNA CRB for providing RNA samples for the project (http://dog-genetics.genouest.org), and the biobank at University of Copenhagen. They are grateful to referring veterinarians, all dog owners and breeders for sample collection and also colleagues from the European LUPA consortium (www.eurolupa.org), especially Anne-Sophie Lequarré and Marilou Ramos for coordinating the program. We also thank FEELnc users and more particularly Kevin Muret, Sandrine Lagarrigue, Mark Cock and Sarah Djebali. Sequencing was performed at the Broad Institute's Genomics Platform. Finally, we are grateful to Rory Johnson and Jacques Nicolas for useful discussions and the GenOuest Bioinformatics core facility (www.genouest.org) for storing sequencing data, hosting the Cani-DNA website and for the use of the bioinformatic cluster to analyze the data.

FUNDING

The 7th PCRD 'Health programs' LUPA consortium [FP7, GA: 201370]; French National Research Agency in the frame of the 'Investing for the Future' program [ANR-11-INBS-0003 to The 'Cani-DNA' biobank, which is part of the CRB-Anim infrastructure] (https://www.crb-anim.fr); European Young Investigator Award from European Science Foundation as well as a Consolidator Award from the European Research Council [to K.L.T.].

Conflict of interest statement. None declared.

REFERENCES

1. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
2. Pervouchine,D.D., Djebali,S., Breschi,A., Davis,C.A., Barja,P.P., Dobin,A., Tanzer,A., Lagarde,J., Zaleski,C., See,L.-H. *et al.* (2015) Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat. Commun.*, **6**, 1–11.
3. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
4. Brown,J.B., Boley,N., Eisman,R., May,G.E., Stoiber,M.H., Duff,M.O., Booth,B.W., Wen,J., Park,S., Suzuki,A.M. *et al.* (2014) Diversity and dynamics of the Drosophila transcriptome. *Nature*, **512**, 393–399.
5. Legeai,F. and Derrien,T. (2015) Identification of long non-coding RNAs in insects genomes. *Curr. Opin. Insect Sci.*, **7**, 37–44.
6. Paytuví Gallart,A., Hermoso Pulido,A., Anzar Martínez de Lagrán,I., Sanseverino,W. and Aiese Cigliano,R. (2016) GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Res.*, **44**, D1161–D1166.
7. Li,L., Eichten,S.R., Shimizu,R., Petsch,K., Yeh,C.-T., Wu,W., Chettoor,A.M., Givan,S.A., Cole,R.A., Fowler,J.E. *et al.* (2014) Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol.*, **15**, R40.
8. Brockdorff,N., Ashworth,A., Kay,G.F., McCabe,V.M., Norris,D.P., Cooper,P.J., Swift,S. and Rastan,S. (1992) 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*, **71**, 515–526.
9. Prensner,J.R. and Chinnaiyan,A.M. (2011) The Emergence of lncRNAs in Cancer Biology. *Cancer Discov.*, **1**, 391–407.
10. Leucci,E., Vendramin,R., Spinazzi,M., Laurette,P., Fiers,M., Wouters,J., Radaelli,E., Eyckerman,S., Leonelli,C., Vanderheyden,K. *et al.* (2016) Melanoma addiction to the long non-coding RNA SAMMSON. *Nature*, **531**, 518–522.
11. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
12. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
13. Haas,B.J., Papanicolaou,A., Yassour,M., Grabherr,M., Blood,P.D., Bowden,J., Couger,M.B., Eccles,D., Li,B., Lieber,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8**, 1494–1512.
14. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.
15. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
16. Wang,L., Park,H.J., Dasari,S., Wang,S., Kocher,J.-P. and Li,W. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74–e74.
17. Li,A., Zhang,J. and Zhou,Z. (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, **15**, 1–10.
18. Sun,L., Luo,H., Bu,D., Zhao,G., Yu,K., Zhang,C., Liu,Y., Chen,R. and Zhao,Y. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.*, **41**, e166.
19. Dewey,C.N. and Pachter,L. (2006) Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Hum. Mol. Genet.*, **15**, R51–R56.
20. Brent,M.R. (2007) How does eukaryotic gene prediction work? *Nat. Biotechnol.*, **25**, 883–885.
21. Blanco,E., Parra,G. and Guigó,R. (2007) Using geneid to identify genes. *Curr. Protoc. Bioinformatics*, **4**, 3.
22. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S., Gil,L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
23. Johnson,R. and Guigo,R. (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*, **20**, 959–976.
24. Zucchelli,S., Fasolo,F., Russo,R., Cimatti,L., Patrucco,L., Takahashi,H., Jones,M.H., Santoro,C., Sblattero,D., Cotella,D. *et al.* (2015) SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells. *Front. Cell. Neurosci.*, **9**, 1720.
25. Tagu,D., Colbourne,J.K. and Nègre,N.N. (2014) Genomic data integration for ecological and evolutionary traits in non-model organisms. *BMC Genomics*, **15**, 1–16.
26. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
27. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
28. Zhao,Y., Li,H., Fang,S., Kang,Y., Wu,W., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.
29. Lequarré,A.-S., Andersson,L., André,C., Fredholm,M., Hitte,C., Leeb,T., Lohi,H., Lindblad-Toh,K. and Georges,M. (2011) LUPA: a European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. *Vet. J.*, **189**, 155–159.
30. Andersson,L., Archibald,A.L., Bottema,C.D., Brauning,R., Burgess,S.C., Burt,D.W., Casas,E., Cheng,H.H., Clarke,L.,

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

Couldrey,C. *et al.* (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.*, **16**, 57.
31. Bussotti,G., Raineri,E., Erb,I., Zytnicki,M., Wilm,A., Beaudoing,E., Bucher,P. and Notredame,C. (2011) BlastR–fast and accurate database searches for non-coding RNAs. *Nucleic Acids Res.*, **39**, 6886–6895.
32. Drezen,E., Rizk,G., Chikhi,R., Deltel,C., Lemaitre,C., Peterlongo,P. and Lavenier,D. (2014) GATB: Genome Assembly & Analysis Tool Box. *Bioinformatics*, **30**, 2959–2961.
33. Claverie,J.M., Sauvaget,I. and Bougueleret,L. (1990) K-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. *Methods Enzymol.*, **183**, 237–252.
34. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**,1915–1927 .
35. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G. *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.
36. Achawanantakun,R., Chen,J., Sun,Y. and Zhang,Y. (2015) LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. *Bioinformatics*, **31**, 3897–3905.
37. Lertampaiporn,S., Thammarongtham,C., Nukoolkit,C., Kaewkamnerdpong,B. and Ruengjitchatchawalya,M. (2014) Identification of non-coding RNAs with a new composite feature in the hybrid random forest ensemble algorithm. *Nucleic Acids Res.*, **42**, e93.
38. Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R. News*, **2**, 18–22.
39. Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.
40. Mattick,J.S. and Rinn,J.L. (2015) Discovery and annotation of long noncoding RNAs. *Nat. Struct. Mol. Biol.*, **22**, 5–7.
41. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
42. Ulitsky,I. (2016) Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Publishing Group*, **17**, 601–614.
43. Duret,L., Chureau,C., Samain,S., Weissenbach,J. and Avner,P. (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, **312**, 1653–1655.
44. Jiang,M., Anderson,J., Gillespie,J. and Mayne,M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 192–211.
45. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
46. Powers,D.M.W. (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2**, 37–63.
47. Askarian-Amiri,M.E., Crawford,J., French,J.D., Smart,C.E., Smith,M.A., Clark,M.B., Ru,K., Mercer,T.R., Thompson,E.R., Lakhani,S.R. *et al.* (2011) SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA*, **17**, 878–891.
48. Djebali,S., Wucher,V., Foissac,S., Hitte,C., Corre,E. and Derrien,T. (2017) Bioinformatics pipeline for transcriptome sequencing analysis. *Methods Mol. Biol.*, **1468**, 201–219.
49. Hoeppner,M.P., Lundquist,A., Pirun,M., Meadows,J.R.S., Zamani,N., Johnson,J., Sundström,G., Cook,A., FitzGerald,M.G., Swofford,R. *et al.* (2014) An Improved Canine Genome and a Comprehensive Catalogue of Coding Genes and Non-Coding Transcripts. *PLoS One*, **9**, e91172.
50. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
51. Lykke-Andersen,S., Chen,Y., Ardal,B.R., Lilje,B., Waage,J., Sandelin,A. and Jensen,T.H. (2014) Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes Dev.*, **28**, 2498–2517.
52. Chen,J., Shishkin,A.A., Zhu,X., Kadri,S., Maza,I., Guttman,M., Hanna,J.H., Regev,A. and Garber,M. (2016) Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol.*, **17**, 19.
53. Sigova,A.A., Mullen,A.C., Molinie,B., Gupta,S., Orlando,D.A., Guenther,M.G., Almada,A.E., Lin,C., Sharp,P.A., Giallourakis,C.C. *et al.* (2013) Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 2876–2881.
54. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
55. Carrieri,C., Cimatti,L., Biagioli,M., Beugnet,A., Zucchelli,S., Fedele,S., Pesce,E., Ferrer,I., Collavin,L., Santoro,C. *et al.* (2012) Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. **491**, 454–457.
56. He,Y., Vogelstein,B., Velculescu,V.E., Papadopoulos,N. and Kinzler,K.W. (2008) The antisense transcriptomes of human cells. *Science*, **322**, 1855–1857.
57. Vaysse,A., Ratnakumar,A., Derrien,T., Axelsson,E., Rosengren Pielberg,G., Sigurdsson,S., Fall,T., Seppala,E.H., Hansen,M.S.T., Lawley,C.T. *et al.* (2011) Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.*, **7**, e1002316.
58. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
59. Herrero,J., Muffato,M., Beal,K., Fitzgerald,S., Gordon,L., Pignatelli,M., Vilella,A.J., Searle,S.M.J., Amode,R., Brent,S. *et al.* (2016) Ensembl comparative genomics resources. *Database (Oxford)*, **2016**, bav096.
60. Hao,Y., Wu,W., Shi,F., Dalmolin,R.J., Yan,M., Tian,F., Chen,X., Chen,G. and Cao,W. (2015) Prediction of long noncoding RNA functions with co-expression network in esophageal squamous cell carcinoma. *BMC Cancer*, **15**, 168.
61. Gutschner,T., Hämmerle,M., Eissmann,M., Hsu,J., Kim,Y., Hung,G., Revenko,A., Arun,G., Stentrup,M., Gross,M. *et al.* (2013) The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.*, **73**, 1180–1189.
62. Hu,G., Tang,Q., Sharma,S., Yu,F., Escobar,T.M., Muljo,S.A., Zhu,J. and Zhao,K. (2013) Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat. Immunol.*, **14**, 1190–1198.
63. LeCun,Y., Bengio,Y. and Hinton,G. (2015) Deep learning. **521**, 436–444.
64. Sharon,D., Tilgner,H., Grubert,F. and Snyder,M. (2013) A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.*, **31**, 1009–1014.
65. Milligan,M.J. and Lipovich,L. (2014) Pseudogene-derived lncRNAs: emerging regulators of gene expression. *Front Genet*, **5**, 476.
66. Johnsson,P., Ackley,A., Vidarsdottir,L., Lui,W.-O., Corcoran,M., Grandér,D. and Morris,K.V. (2013) A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nat. Struct. Mol. Biol.*, **20**, 440–446.
67. Ørom,U.A. and Shiekhattar,R. (2013) Long noncoding RNAs usher in a new era in the biology of enhancers. **154**, 1190–1193.
68. Ren,B. (2010) Transcription: enhancers make non-coding RNA. **465**, 173–174.
69. Hedges,S.B., Marin,J., Suleski,M., Paymer,M. and Kumar,S. (2015) Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.*, **32**, 835–845.

Downloaded from http://nar.oxfordjournals.org/ at BibliothÃ¨que de l'IRMAR on January 10, 2017

UNIVERSITE BRETAGNE LOIRE / BIOLOGIE SANTE

UNIVERSITÉ DE RENNES 1

Titre : Analyse intégrative des ARN longs non-codants chez le chien et leurs implications dans le mélanome oral canin, modèle des mélanomes humains

Mots clés : lncRNA, chien, RNA-seq, transcriptome, génomique comparative, élément transposable, cancers

Résumé : Les ARN longs non-codants (lncRNAs) constituent une famille d'ARN hétérogènes qui jouent un rôle majeur dans de nombreux cancers et notamment dans les mélanomes. Le chien est un modèle naturel et spontané pour l'analyse génétique comparée des cancers et, l'annotation du génome canin a récemment été enrichie avec l'identification de plus 10 000 lncRNAs. Afin de réaliser des prédictions fonctionnelles bioinformatiques des lncRNAs, nous avons caractérisé les profils d'expression des lncRNAs canins à partir de 26 tissus distincts. Nous avons défini la spécificité tissulaire de l'expression des lncRNAs et inféré leur fonctionnalité potentielle par des analyses de génomique et de transcriptomique comparatives avec des données humaines issues du projet ENCODE (ENCyclopedia Of DNA Elements). Comme chez l'homme et la souris, une grande proportion de lncRNAs canins (44 %) est exprimée de manière spécifique au sein d'un tissu. Par une approche de génomique comparative, nous avons identifié plus de 900 lncRNAs orthologues entre l'homme et le chien et pour 26 % d'entre eux, des patrons d'expression entre tissus significativement conservés ($p < 0.05$).

Dans le cadre de l'étude des mélanomes canins, nous avons analysé les données de RNA-seq de 52 échantillons tumeurs/contrôles de mélanomes oraux. Nous avons identifié plus de 750 lncRNAs différentiellement exprimés entre la tumeur et le contrôle (FDR < 0.01), dont plus de 100 conservés avec l'homme. Ces lncRNAs constituent de bons candidats pour étudier la régulation de la progression tumorale des mélanomes chez le chien et pourront être évalués pour leurs potentiels diagnostic et thérapeutique en médecine humaine et vétérinaire.

Title: Integrative analysis of long non-coding RNAs in dogs and their implications in canine oral melanoma, human melanoma model

Keywords: lncRNA, dog, RNA-seq, transcriptome, comparative genomics, transposable element, cancers

Abstract: Long non-coding RNAs (lncRNAs) are a family of heterogeneous RNAs that play a major role in many cancers, particularly in melanomas. The dog is a natural and spontaneous model for the comparative genetic analysis of cancers and, the annotation of the canine genome has recently been enriched with the identification of over 10,000 lncRNAs. In order to perform functional bioinformatic predictions of lncRNAs, we have characterized the expression patterns of canine lncRNAs from 26 distinct tissues representative of the major functions of the organism. We defined the tissue specificity of lncRNAs expression and inferred their potential functionality by comparative genomic and transcriptomic analyses with human data from the ENCODE project (ENCyclopedia Of DNA Elements). As in humans and mice, we show that a large proportion of canine lncRNAs (44%) are expressed specifically within a tissue. Using a comparative genomic approach, we have identified more than 900 orthologue lncRNAs between humans and dogs, and we show that for 26% of them, tissue expression patterns are also significantly conserved ($p < 0.05$).

In the study of canine melanomas, we investigated the lncRNAs from RNA-seq data from 52 tumour/control samples of oral melanoma. We identified more than 750 lncRNAs differentially expressed between tumour and control (FDR < 0.01), of which more than 100 were conserved with humans. These lncRNAs are good candidates to study the regulation of tumour progression of melanomas in dogs and can be evaluated for their diagnostic and therapeutic potential in human and veterinary medicine.