Utilisation des séquences de génome complet pour l'identification de mutations délétères responsables d'anomalies génétiques récessives chez le bovin.

Pauline Michot

NNT : 2017SACLA011

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT (AGROPARISTECH)

ÉCOLE DOCTORALE N°581
Agriculture, alimentation, biologie, environnement et santé

Spécialité de doctorat : Génétique Animale

Par

Melle Pauline Michot

Utilisation des séquences de génome complet pour l'identification de mutations délétères responsables d'anomalies génétiques récessives chez le bovin

Directeur de thèse : Didier Boichard & Aurélien Capitan

Thèse présentée et soutenue à Paris, le 29 Mars 2017 :

Composition du Jury :

M. Etienne VERRIER	Professeur, AgroParisTech, Paris	Président
Mme Carole CHARLIER.	Maître de recherche du FNRS, Université de Liège	Rapporteur
Mme Jeanne AMIEL	PUPH, Institut Imagine Hôpital Necker, Paris	Rapporteur
Mme Catherine ANDRE	Chargée de recherche, CNRS, Université de Rennes 1	Examinateur
M. Laurent SCHIBLER	Responsable Développement & Innovation., ALLICE, Paris	Examinateur
M. Didier BOICHARD	Directeur de recherche, INRA, Jouy en Josas	Examinateur

Remerciements

Tous ceux qui ont suivi mon parcours de thèse savent à quel point je suis soulagée de pouvoir enfin écrire les dernières lignes de ce document. Il achève un peu plus de trois ans de travaux au sein de l'INRA de Jouy-en-Josas, parfois difficiles et moralement éprouvants sur la fin, mais sur une thématique de recherche toujours aussi passionnante. Aussi, je souhaite adresser mes plus profonds et sincères remerciements à toutes les personnes rencontrées au cours de ma thèse, qui ont contribué à mes travaux et fait de ces trois années une expérience professionnelle et humaine extrêmement enrichissante.

Tout d'abord, je remercie mes directeurs de thèse, Aurélien Capitan et Didier Boichard, qui m'ont donné la chance de pouvoir entreprendre cette thèse et sans qui ce document n'aurait pu voir le jour. Un profond merci à vous deux pour vos enseignements, le partage de vos connaissances scientifiques, de recherche et de votre passion pour la génétique. Merci pour votre confiance et votre soutien tout au long de ma thèse, malgré quelques moments de doute. J'ai eu beaucoup de chance de la réaliser sous votre direction.

A Xavier David, directeur d'ALLICE, et Laurent Schibler, responsable R&D ; j'adresse mes très sincères remerciements pour la confiance et l'accueil que vous m'avez accordés dans la réalisation de ce projet de thèse au sein d'ALLICE. Je remercie aussi l'ensemble du personnel de l'entreprise pour les échanges toujours chaleureux.

Je remercie l'ANRT et APIS-GENE qui ont co-financé ma thèse dans le cadre d'un contrat CIFRE, ainsi que l'Agence Nationale de la Recherche et APIS-GENE pour le financement du projet BOVANO.

Merci à l'ensemble de l'équipe G2B et personnels de l'INRA avec qui j'ai pu travailler ou simplement échanger au cours de ces trois années. Plus particulièrement à Sébastien Fritz, mon directeur de stage de fin d'étude, merci encore pour tous vos conseils et vos encouragements ; à Cécile Grohs, véritable « Maman labo », merci pour votre implication dans les projets de cette thèse, votre aide, votre écoute et conseils à la paillasse, sans oublier non plus votre amitié et vos encouragements.

A tous les collègues, doctorants et stagiaires de l'étage et/ou du rez-de-chaussée du 211 - en particulier Chris, Romain, Laure, Thierry, Alexis, Iola, Gabriel – merci pour les échanges et discussions au sein de l'équipe, merci pour votre amitié, les conversations animées des pauses café, les moments « geek » et batailles de nerfs, les blagues vraiment pas drôles, les afterworks Baradoz… Merci pour tous ces bons moments qui vont me manquer. Je vous souhaite à tous une très bonne poursuite dans vos différents projets. Mentions spéciales à mes trois co-bureaux successifs, Marine, Alexis et enfin Rabia pour quelques semaines de rédaction pas les plus drôles pour moi. Merci à vous pour ce partage toujours dans une bonne ambiance.

Une pensée à tous mes amis de Bourgogne, de prépa, de l'agro, des USA qui n'ont jamais très bien compris mon sujet de recherche ou bien même l'utilité d'une thèse, mais toujours présents pour décompresser.

Pour finir, un immense merci à l'ensemble de mes proches qui m'ont soutenue. A mes parents, anciens éleveurs de Charolaises, qui m'ont transmis la passion de l'élevage, à mon grand-frère, exemple et soutien depuis toujours; à ma belle-sœur pour ses conseils, et mon copain qui m'a supportée et soutenue dans ces mois de rédaction ; merci à vous tous du fond du cœur pour votre patience, vos conseils et votre force.

A nouveau, à toutes les personnes citées ci-dessus ou non, merci.

Mots clés : génomique, mutations délétères, anomalies, bovins

Résumé

L'effectif génétique réduit des races bovines entraîne une augmentation de consanguinité de l'ordre de 1% par génération et une forte dérive génétique. Cette évolution favorise l'émergence régulière d'anomalies génétiques récessives dans les populations, qu'elles soient de races laitières ou allaitantes. En France, l'Observatoire National des Anomalies Bovines (ONAB) a été créé dans le but de détecter et contrôler ces anomalies émergentes. Cependant, la détection par les observatoires sous-entend d'une part une diffusion large de l'allèle défavorable dans la population et d'autre part que l'anomalie présente un phénotype avec un tableau clinique spécifique permettant sa déclaration à l'ONAB. L'impact des anomalies génétiques est donc encore largement sous-estimé. Toutefois, les développements des technologies de génotypage et de séquençage de génomes, associés à l'ensemble des informations disponibles permettent une détection efficace des mutations causales. Ainsi, l'objectif de cette thèse a été d'utiliser l'ensemble des données disponibles (phénotypes, génotypes, séquences, annotations fonctionnelles…) pour identifier et valider des mutations délétères ségrégant dans races bovines laitières et allaitantes françaises.

Nous avons exploré différentes stratégies classiques - cartographies par homozygotie, haplotypes en déficit en homozygotes - qui gagnent en efficacité grâce aux données de séquence de génome complet (WGS). Nous avons également mis en place des approches alternatives de génétique inverse, basées sur l'exploitation des données WGS françaises et du consortium « 1000 bull genomes ». Les travaux réalisés ont permis d'identifier les mutations causales associées à deux syndromes récessifs rapportés à l'ONAB : l'épidermolyse bulleuse jonctionnelle en race Charolaise (ITGB4, chr19: g.56488278_56493087del) et l'épilepsie idiopathique (MTCL1, chr24:g.41661691G>A) récemment émergée dans la race Parthenaise. Nous avons également démontré la forte association entre l'haplotype MH1 et un polymorphisme affectant le gène PFAS (chr19:g.28511199C>T ; p.R1205C). Par les stratégies de génétique inverse, nous avons également identifié une mutation probablement responsable de mortalité embryonnaire en race Normande affectant le gène CAD (chr11 :g72399397, p.Y452C) ainsi qu'une mutation affectant le gène RP1 (chr14:g.23995411_23995412insA, p. R791KfsX13) responsable d'une dégénérescence progressive de la rétine qui ségrége à forte fréquence en race Normande mais aussi dans d'autres races bovines européennes. Ces études encore en cours, fournissent également un inventaire des variants génétiques potentiellement délétères dont la caractérisation de l'effet sur le phénotype pourra être explorée.

Enfin, l'identification de ces anomalies et des mutations délétères responsables ont abouti à la mise à disposition de tests de diagnostic efficaces pour permettre une contre sélection raisonnée de ces variants délétères dans les populations bovines.

Keywords: genomic, deleterious mutations, anomalies, bovine

Abstract

The reduced genetic size of cattle breeds leads to an 1% increase in inbreeding per generation and a strong genetic drift. This evolution favors regular emergences of recessive genetic abnormalities in dairy and beef cattle breeds. In France, the National Observatory of Bovine Anomalies (ONAB) was created with the aim to detect and control these new emergences. However, detection by the observatories implies a broad diffusion of the deleterious allele in the population as well as a phenotype with a specific clinical features, which allows reporting of the anomaly to the ONAB. Therefore the impact of genetic anomalies is still largely underestimated. However, development of genotyping and genome sequencing technologies, coupled with all available information, allows effective detection of causal mutations. Thus, the aim of this thesis was to use all the available data (phenotypes, genotypes, sequences and functional annotations) to identify and validate deleterious mutations segregating in French dairy and beef cattle breeds.

We explored various classical strategies, such as homozygosity mapping and search for haplotypes displaying a deficit in homozygotes, which gained efficiency with whole genome sequence (WGS) data. We also implemented alternative reverse genetics strategies, based on data mining of French and "1000 bull genomes" consortium WGS data. In these different studies, we identified the causal mutations associated with two recessive syndromes: epidermolysis bullosa junctional in Charolais breed (ITGB4, chr19: g.56488278_56493087del) and idiopathic epilepsy (MTCL1, chr24:g.41661691G>A) recently emerged in the Parthenaise breed. We also demonstrated a strong association between the embryonic lethal mutation MH1 and a polymorphism affecting the *PFAS* gene (chr19: g.28511199C> T ; p.R1205C) in Montbeliarde cattle. With reverse genetic strategies, we identified another mutation in the CAD gene (chr11: g72399397, p.Y452C), likely responsible for embryonic mortality in Normande cattle and a frameshift mutation in *RP1* (chr14:g.23995411_23995412insA, p. R791KfsX13) responsible for progressive retinal degeneration segregating with a high frequency in the Normand breed and also in other European cattle breeds. These studies, still in progress, provide an inventory of potentially deleterious genetic variants, the characterization of them could be explored in the future.

At last, identification of mutations responsible for these genetic abnormalities provides effective diagnostic tests and allows judicious counter selection of these variants in French beef and dairy cattle populations.

Table des matières

Liste des figures

Liste des tableaux

Liste des Abréviations

ADN : Acide DésoxyriboNucléique
ARN : Acide RiboNucléique
BAM : Binary Alignment Map
BLAD : Bovine Leucocyte Adhésion Deficiency
BWA : Burrows-Wheeler Aligner
CAD : Carbamoyl-Phosphate Synthetase 2, Aspartate Transcarbamylase, And Dihydroorotase
CVM : Complex Vertebral Malformations
EBJ : Epidermolyse Bulleuse Jonctionnelle
ExAC : Human Exome Aggregation Consortium
FANC1 : Fanconi Anemia Complementation Group I, responsable de l'anomalie Brachyspina
FCEL: France Conseil en Elevage
FGE : France Génétique Elevage (interprofession génétique des ruminants)
G2B : Equipe de Génétique et Génomique Bovine
GABI : Unité Mixte de Recherche « Génétique Animale et Biologie Intégrative »
GATK : Genome Analysis Tool Kit
HW Loi de Hardy Weinberg
IBD : Identity By Descent
IDELE : Institut de l'Elevage
IGV : Integrative Genome Viewer
InDel : Insertion / Délétion
ITGB4 : Integrin Beta 4
LOF : Loss Of Function
MARK2 : Microtubule Affinity Regulating Kinase 2
MGI : Mouse Genome Informatics, base de données du Jackson Laboratory
MH1, MH2 : Montbeliarde Haplotype 1, Montbeliarde Haplotype 2
MTCL1 : Microtubule Crosslinking Factor 1
NCBI : National Center for Biotechnology Information
NGS : New Generation Sequencing
NRR56 : Taux de Non Retour à 56 jours
OMIA : Online Mendelian Inheritance in Animals
OMIM : Online Mendelian Inheritance in Man
ONAB : Observatoire National des Anomalies Bovines
OS : Organisme de Sélection
PFAS : Phosphoribosylformylglycinamidine synthase
QTL : Quantitative Trait Locus
RFLP : Restriction Fragment Length Polymorphism
RP1 : Retinitis Pigmentosa 1
SHBG : Sex-Hormone Binding Globulin
SLC37A2 : Solute Carrier Family 37 A2
SNGTV : Société Nationale des Groupements Techniques Vétérinaires

SNP : Single Nucleotide Polymorphism
UCSC : University of California Santa Cruz
UMT 3G : Unité Mixte Technologique “Gestion Génétique et Génomique des Populations Bovines”
UNCEIA (devenue ALLICE) : Union Nationale des Centres d’Elevage et d’Insémination Animale
VEP : Variant Effect Predictor
WGS : Whole Genome Sequence

CHAPITRE 1. INTRODUCTION BIBLIOGRAPHIQUE : LES ANOMALIES GENETIQUES RECESSIVES CHEZ LE BOVIN

Ce chapitre constitue une introduction à mes travaux de thèse, avec dans un premier temps une rapide description de l'élevage bovin, l'organisation de la sélection des races et la structure génétique des populations. Je présenterai ensuite une introduction aux anomalies génétiques chez les bovins et les mesures de gestion mises en place, puis les évolutions technologiques en génétique moléculaire et les sources d'information disponibles sur lesquelles les travaux sur les nouvelles anomalies peuvent s'appuyer aujourd'hui. Enfin, je présenterai rapidement les méthodologies utilisées dans la recherche des mutations causales à déterminisme supposé récessif et employées au cours de ma thèse.

I. Les anomalies génétiques chez le bovin : comprendre les émergences

A. Organisation de l'élevage bovin en France

1. L'élevage bovin français : quelques chiffres

Premier à l'échelle européenne, le cheptel bovin français compte 19,4 millions d'individus sur le territoire répartis pour 42% de l'effectif total en systèmes de production laitière et 58% en système allaitant (Agreste). Ces effectifs représentent 3,7 millions de vaches laitières et 4,1 millions de vaches allaitantes adultes recensées dans les fermes en 2015 (France Génétique Elevage (FGE)). La population est très majoritairement structurée en races, le croisement étant minoritaire et limité pour l'essentiel à la production d'animaux de boucherie non reproducteurs. Une race est une population relativement homogène, avec un standard défini, et conduite, sauf exception, de façon fermée. La population française est répartie en huit races laitières et neuf races allaitantes principales, auxquelles s'ajoutent 4 races à petits effectifs (ie. Bazadaise, Bleu du Nord, Rouge Flamande et Vosgienne) et 15 races à très petits effectifs en processus de conservation. Pour les races laitières, la Holstein représente à elle seule 66% de l'effectif avec 2,47 millions de vaches adultes. Viennent ensuite les races Montbéliarde et Normande qui comptent respectivement 648 (17%) et 357 milliers (9%) de vaches. Toutes trois ont un rayonnement national, voire international dans le cas de la race Holstein, et constituent plus de 95% des effectifs laitiers. En comparaison, cinq races avec des effectifs compris entre 13 et 48 milliers de vaches et une aire de répartition limitée rassemblent seulement 3,4% de l'effectif total. Il s'agit de deux races locales des Alpes du Nord, l'Abondance et la Tarentaise, de deux noyaux français de races étrangères, la Simmental Française et la Brune, et de la race Pie Rouge. On compte également une petite fraction de

vaches de races à petits effectifs ou de vaches croisées. La situation est similaire en production allaitante où trois races nationales, la Charolaise (1,5 millions de vaches), la Limousine (1 million) et la Blonde d'Aquitaine (520 milliers) représentent à elles seules 78% du cheptel. Viennent suite cinq races locales rustiques (Salers, Aubrac, Gasconne) ou à orientation viande (Rouge des Prés, Parthenaise).

2. Gestion des races et organisation des programmes de sélection

Toutes les races, à l'exception de celles en conservation, font l'objet d'une sélection, organisée selon un programme de sélection. Tous les animaux disposent d'un identifiant unique. La sélection sur la voie mère fille est peu intense, car la précision de l'estimation de la valeur génétique des femelles d'une part, leur prolificité et donc l'intensité de sélection d'autre part, sont limitées. L'essentiel de la sélection passe par la voie mâle, avec les trois voies concernées : mère fils, père fils, père fille. La race est structurée en deux grands étages au minimum, les sélectionneurs et les utilisateurs. Les sélectionneurs produisent les reproducteurs utilisés à la fois chez les sélectionneurs et chez les utilisateurs. Jusqu'à l'avènement récent de la sélection génomique, la sélection des reproducteurs a reposé sur les performances phénotypiques des candidats et de leurs apparentés. Dans l'étage de sélection, des procédures sont donc mises en place pour mesurer les phénotypes et connaitre les pedigrees.

Les populations laitières et allaitantes présentent plusieurs différences. Tout d'abord, elles ne sont pas sélectionnées sur les mêmes caractères et le contrôle de performances associé est donc adapté à chaque filière. Par ailleurs, si l'insémination artificielle est très majoritairement utilisée en système laitier, elle est minoritaire en système allaitant où la monte naturelle est le mode de reproduction prédominant. Le besoin de taureaux est donc sensiblement différent. Les taureaux d'insémination sont relativement peu nombreux, largement diffusés et doivent être fortement sélectionnés et précisément évalués, ce qui nécessite un testage sur descendance. Les taureaux de monte naturelle sont nombreux, ils ont moins de descendants, leur pression de sélection ainsi que la précision d'évaluation sont modérées et cette précision est obtenue à partir d'informations sur ascendance et de performances propres. Les taureaux laitiers sont évalués sur des caractères pour la plupart mesurables uniquement sur les femelles et cette évaluation sur descendance est réalisée en ferme. Les taureaux allaitants sont évalués sur des caractères variés, mesurés en ferme ou en station. Le processus de sélection comporte généralement plusieurs étapes : choix du père et de la mère parmi les meilleurs parents disponibles, puis du taureau au cours de plusieurs étapes successives incluant une évaluation sur performance propre ou sur descendance.

La diffusion du progrès génétique a lieu majoritairement par la voie mâle, donc par l'utilisation des taureaux sélectionnés, soit par insémination artificielle, soit par monte naturelle à la suite de la vente du reproducteur.

B. Une structuration de la population favorable à l'émergence d'anomalies génétiques

1. Des populations à effectifs génétiques faibles

Depuis leur domestication, les bovins ont fait l'objet d'une sélection phénotypique pour divers caractères en fonction de l'évolution de leur utilisation et de nos besoins (docilité, adaptation à un environnement, aptitudes au travail, production laitière et bouchère ...). A partir de la fin du 18ème siècle et majoritairement au 19ème siècle, les races ont été constituées en rassemblant des animaux aux aptitudes comparables issus de la même région, en les faisant se reproduire entre eux et en standardisant progressivement leur apparence, en particulier la couleur de leur robe. A l'échelle de l'espèce, la diversité des phénotypes entre races reflète de manière intrinsèque une grande variabilité génétique et la sélection préférentielle de gènes et mutations (Andersson, 2001). Cependant au sein d'une race, si les effectifs sont parfois grands, en particulier dans les races nationales, la diversité génétique est modérée à faible et se réduit peu à peu.

Tableau 1: Effectifs génétiques (Ne) des principales races bovines laitières et allaitantes (période 2012/2015).

Race laitières	Ne	Race allaitantes	Ne
Abondance	51	**Aubrac**	108
Brune	93	**Blonde d'Aquitaine**	70
Montbéliarde	75	**Charolaise**	175
Normande	84	**Gasconne**	89
Prim'Holstein	96	**Limousine**	149
Simmental	139	**Parthenaise**	81
Tarentaise	65	**Rouge des Prés**	73
		Salers	112

Source : Institut de l'élevage VARUME résultats 2016

Une méthode de suivi de la variabilité génétique des populations est l'utilisation de différents indicateurs résultant de l'analyse des pedigrees et de la probabilité d'origine de gènes (Boichard *et al.*, 1997): le nombre d'ancêtres efficaces (*Ae*), le nombre d'ancêtres principaux contributeurs représentant 50 % de l'origine des gènes de la population (*N50*) et l'estimation de l'effectif génétique ou taille efficace (*Ne*). Depuis 1996, les bilans de variabilité génétique des principales populations bovines françaises ont été régulièrement établis (Boichard *et al.*, 1996 ; Moureaux *et al.*, 2000 ; Mattalia *et al.*, 2006 ; Bouquet *et al.*, 2009 ; Danchin *et al.*, 2012) et un bilan annuel dans toutes les races bovines est maintenant publié par l'IDELE au travers de l'observatoire de variabilité génétique des ruminants (VARUME) (Danchin-Burge *et al.*, 2014). De 1988 à 2007, presque toutes les races laitières ont connu une réduction forte des

valeurs *Ae* et *N*50. Pour les trois races laitières nationales Normande, Montbéliarde et Holstein, le *N50* a diminué de 53%, 65% et 72%. Moins de 10 ancêtres (respectivement 7 en Montbéliarde et 8 en Normande et Holstein) contribuent à 50% des gènes de la population femelle de chaque race, tandis que l'effectif génétique intra race varie de 50 à 100 (Tableau 1), indépendamment de l'effectif réel de femelles. En races allaitantes, les effectifs génétiques sont plus élevés qu'en races laitières du fait de l'utilisation d'un bien plus grand nombre de reproducteurs. Ces valeurs restent cependant faibles, si on les compare à l'homme par exemple (600 à 10000 selon les estimations, Park, 2011).

La structuration particulière de la population bovine est le résultat d'une part de l'échantillonnage initial restreint à l'origine de chaque race, et d'autre part de goulets d'étranglement sévères appliqués par les évolutions des pratiques d'élevage et l'intensification de la production dans la seconde moitié du 20ème siècle. Le développement des biotechnologies de la reproduction, en particulier de l'insémination animale réalisée « en frais » à partir de 1950 puis l'invention des paillettes et le développement de la congélation de la semence au début des années 1970 (Cassou, 1968), ont permis la diffusion rapide à très large échelle des meilleurs taureaux reproducteurs. Le succès de l'IA a également favorisé le développement de la sélection qui, dans la recherche constante d'un meilleur gain génétique, s'est fortement intensifiée au début des années 1990. Ceci s'est traduit par l'application d'une très forte pression de sélection sur les mâles à l'entrée en station, un faible nombre de taureaux diffusés par année (quelques dizaines par race), ainsi que l'utilisation souvent excessive des meilleurs taureaux, qui deviennent rapidement des contributeurs majeurs en termes de patrimoine génétique à l'échelle de leur race (Mattalia *et al.*,2006). Dans les années 2000, une prise de conscience a permis de limiter ces excès (Verrier *et al.*, 2005) et la sélection génomique a été l'occasion de diffuser moins largement les reproducteurs, eux-mêmes en plus grand nombre.

2. Conséquences : augmentation de la consanguinité et dérive génétique

La réduction du nombre d'ancêtres contributeurs et l'effectif génétique réduit des races a deux conséquences principales. La première est une augmentation du taux de consanguinité avec le temps. Dans les races laitières, elle s'est élevée au rythme de 0,5 à 1 % par génération sur la période de 1996 à 2006 (Boichard *et al.*, 1996 ; Mattalia *et al.*,2006 ; Danchin *et al.*, 2012).

Aujourd'hui, par rapport à 1960, les taux de consanguinité se situent entre 4,2 et 4,7% dans les races laitières et entre 0,8 et 1,4 % dans les races allaitantes qui bénéficient encore d'un effectif génétique plus grand (du fait de la présence encore importante d'un noyau de monte naturelle qui a maintenu un grand nombre de reproducteurs). L'accroissement de consanguinité observé dans les populations traduit directement le taux de perte de variabilité génétique et l'homogénéisation progressive des génomes.

Concernant plus particulièrement les anomalies, la consanguinité est responsable d'une augmentation du taux d'homozygotie du génome chez chaque individu et donc de l'expression des anomalies récessives. Par ailleurs, le faible effectif génétique induit une forte dérive génétique : les fréquences

alléliques au sein de la population varient rapidement du fait de la diffusion massive et privilégiée d'un faible nombre de reproducteurs ou d'origines. Un variant délétère initialement rare peut voir sa fréquence augmenter rapidement en une ou deux générations, préparant ainsi une future émergence.

En conclusion, cette structuration par races à effectif génétique faible, fonctionnant comme des isolats génétiques, crée des conditions propices à l'émergence des anomalies génétiques héréditaires.

3. L'origine des anomalies génétiques

La réplication de l'ADN est source de mutations. Les mutations sont la source initiale de la variabilité génétique. Chez l'homme, ce taux de mutation est de l'ordre de 1.1 à 1.7x10-8 par nucléotide et par génération, ce qui fait, compte tenu de la taille de notre génome, que chaque individu porte de 50 à 100 mutations *de novo*. Des valeurs similaires ont été rapportées chez le bovin (Harland *et al.,* 2016). La plupart des mutations sont neutres et affectent des régions non codantes du génome. Une faible proportion d'entre elles, affectant des régions codantes ou régulatrices de l'expression des gènes peuvent avoir un effet phénotypique favorable ou défavorable. Lorsque l'effet est très délétère à l'échelle de l'organisme, on parle d'anomalie. Parmi les dizaines de millions de variants décrits, en particulier à partir des travaux de séquençage, quelques milliers sans doute peuvent être à l'origine d'anomalies génétiques récessives.

Une mutation produit un variant à très faible fréquence, puisqu'un seul des 2N chromosomes de la population est porteur, N étant l'effectif de la population. Ce variant est ensuite soumis à diverses forces évolutives qui vont faire varier sa fréquence. Pour la grande majorité d'en eux, ces variants disparaissent en une ou quelques générations, faute de transmission à des descendants.

Parmi les forces évolutives auxquelles sont soumises ces mutations, la dérive occupe une place particulière dans nos populations. La dérive est la variation aléatoire de fréquence d'une génération à l'autre, liée à l'échantillonnage des reproducteurs et le nombre fini de descendants par reproducteur. Dans les races bovines, c'est souvent la raison principale de l'émergence de nouvelles anomalies.

En effet, quand un reproducteur a une contribution importante à sa race, il provoque également une augmentation de fréquence des allèles qu'il porte responsables d'anomalies. Les années 1990 et 2000 ont été marquées par l'émergence successive de plusieurs anomalies génétiques récessives, dont les mutations causales ont atteint des fréquences alléliques non négligeables, et touchaient de ce fait de nombreuses naissances par an au moment de leur découverte. Dans la race Holstein, ce sont par exemple les émergences des anomalies BLAD (*Bovine Leucocyte Adhésion Deficiency,* Shuster *et al.,* 1992) et CVM (*Complex Vertebral Malformations,* Thomsen *et al.,* 2006) diffusées massivement par les taureaux d'IA Bell et Star, largement utilisés à l'échelle mondiale au cours des années 1980 et 1990. Ainsi, la fréquence de l'anomalie BLAD atteignait près de 15% de porteurs dans la population de taureaux aux Etats-Unis et 20 % dans les taureaux français au moment du pic d'émergence des cas (Schuster *et al.,*

1992 ; Boichard *et al.,* 1994). Pour CVM, les fréquences ont été estimées à près de 31% de porteurs parmi les taureaux d'élite Danois (Thomsen *et al.*, 2006).

La sélection, naturelle ou artificielle, est une seconde force évolutive responsable de la fluctuation des fréquences alléliques. Ainsi, par exemple, les mutations à effet dominant qui réduisent la probabilité des porteurs de survivre et de se reproduire sont généralement éliminées rapidement, parfois en une génération. Parfois, ce processus peut être ralenti, comme pour les mutations dominantes à pénétrance incomplète, ou associées à un mosaïcisme somatique ou germinal. Au contraire, un allèle favorable et dominant tend à diffuser rapidement. La pression de sélection sur les anomalies récessives est généralement faible, car la plupart des individus porteurs sont hétérozygotes et l'anomalie se comporte chez eux comme un variant neutre. Dans ce cas, la contre-sélection naturelle est généralement moins efficace que la dérive. Toutefois, il existe des situations où la sélection peut conduire à faire augmenter rapidement la fréquence d'un allèle récessif responsable d'une anomalie. Dans le cadre d'une sélection équilibrante, l'hétérozygote présente un avantage sur les deux génotypes homozygotes. Plusieurs exemples ont été rapportés en race Blanc Bleu Belge ces dernières années (par exemple Fasquelle *et al.*, 2009). La sélection peut également être indirecte, lorsqu'une mutation récessive délétère est associée à un allèle favorable pour un QTL de production situé sur le même brin chromosomique et est sélectionné en même temps que ce dernier. On parle alors d'effet d' « auto-stop ». Dans ces deux derniers cas, la fréquence des mutations et la prévalence des anomalies concernées peuvent atteindre un niveau très élevé, rarement atteint en cas de dérive simple.

C. Des dispositifs d'hérédo-surveillance pour repérer les anomalies : les observatoires nationaux

1. L'intérêt des dispositifs pour repérer les anomalies

Pour enrayer l'émergence d'une anomalie, il faut être capable de détecter les cas précocement. Or ceux-ci sont relativement rares (2,5 homozygotes pour 1000 dans le cas classique d'une mutation récessive avec une fréquence de 5% dans la population), dispersés dans beaucoup d'élevages, au sein desquels, le plus souvent, un seul cas est apparu. Il est alors fréquent que le cas isolé soit passé, volontairement ou non, sous silence. Par ailleurs, les systèmes usuels de collecte de phénotypes n'ont pas été conçus pour l'enregistrement des anomalies. On dispose tout au plus de l'information de mortalité, ainsi que des informations de reproduction et fertilité, qui peuvent donner des indices pour les anomalies fréquentes, mais qui sont peu adaptées aux nouvelles émergences. L'expérience a montré que le système français a été incapable de détecter les émergences pourtant assez marquées du BLAD et du CVM, en 1992 et en 2000. Pourtant, les anomalies coûtent cher, et tout particulièrement pour les sélectionneurs, car la valeur d'un reproducteur est sensiblement réduite s'il se révèle porteur d'une anomalie. Par ailleurs, le coût en termes d'image auprès de la société est important, car l'anomalie peut être identifiée comme une atteinte

au bien-être et associée à la sélection. Il est donc essentiel de disposer d'un système de repérage et de description des cas. Dans la mesure où les cas sont souvent isolés, cet observatoire n'a une bonne efficacité que s'il est centralisé.

Les travaux sur les anomalies génétiques sont anciens. Il convient de citer notamment H.W. Leipold et K. Huston (Kansas State University) qui ont joué un rôle précurseur au niveau international, tandis qu'à l'INRA, J.J. Lauvergne a réalisé les premières études sur ce sujet, dans différentes espèces animales d'élevage.

Le Danemark a été précurseur dans la mise en place dès 1989 d'un observatoire systématique, associant un grand réseau de vétérinaires dans le Danish Bovine Genetic Disease Programme (Agerholm *et al.*, 1993), avec des résultats remarquables. En Belgique, l'Université de Liège a également mis en place un réseau de surveillance en race Blanc Bleu Belge, également sous la forme d'un réseau de plus de 200 vétérinaires animé par A. Sartelet. Cette équipe très active a découvert plusieurs anomalies dont celles responsables de nanisme (Sartelet *et al.*, 2012), du syndrome de queue tordue (Fasquelle *et al.*, 2009), ou de l'hamartome (Sartelet *et al.*, 2014). Dans les autres pays, les dispositifs équivalents sont le plus souvent des initiatives de facultés vétérinaires, comme en Allemagne, l'équipe d'O. Distl, à l'Université vétérinaire de Hanovre, ou en Suisse, à l'université de Berne, l'équipe de C. Drögemüller.

2. Présentation du système français : L'Observatoire National des Anomalies Bovines

Après le constat que les anomalies BLAD puis CVM n'avaient pas été détectées en France, le département de Génétique Animale de l'INRA a été à l'origine de la création de l'Observatoire National des Anomalies Bovines (ONAB) en 2002. L'objectif était de rassembler l'ensemble des partenaires intéressés et susceptibles d'observer et de faire remonter des cas, en particulier les inséminateurs, les agents de contrôle de performances et les vétérinaires. L'ONAB a donc regroupé les organismes et fédérations professionnels comme l'Institut de l'Elevage, l'UNCEIA (devenue ALLICE), Races de France, la Société Nationale des Groupements Techniques Vétérinaires (SNGTV) et France Conseil en Elevage (FCEL), sous l'égide du Ministère chargé de l'agriculture (Ducos *et al.*, 2003). L'animation de l'ONAB a été initialement confiée au Laboratoire de Cytogénétique de l'Ecole Vétérinaire de Toulouse, déjà en charge du diagnostic pour les translocations chromosomiques. Elle a ensuite été assurée conjointement par l'unité GABI à Jouy-en-Josas et l'Institut de l'Elevage.

L'ONAB est une structure informelle dont les règles de fonctionnement sont régies par une charte. Aux membres historiques de 2002 se sont ajoutés GDS France (fédération nationale des Groupements de Défense Sanitaire) et les 4 écoles nationales vétérinaires. Depuis 2014, des financements ont été obtenus pour soutenir le fonctionnement de l'observatoire : d'une part, France Génétique Elevage apporte un soutien à l'ONAB, d'autre part le projet de recherche BOVANO financé par l'Agence Nationale pour la Recherche et Apis-Gene pour une durée de 5 ans permet de mettre en œuvre les actions en amont et

en aval de l'ONAB pour étudier les anomalies détectées. La structure, les objectifs, les méthodes, l'activité et les résultats de l'ONAB sont présentés sur son site web *http://www.onab.fr*.

Les partenaires définissent collectivement les actions de l'observatoire au cours de réunions régulières du comité de pilotage. Les objectifs principaux de l'ONAB sont de : (1) détecter précocement l'émergence d'anomalies ; (2) approfondir les connaissances scientifiques en matière d'étiologie et d'épidémiologie des anomalies congénitales ; (3) aider à l'éradication raisonnée des anomalies sur le terrain. Pour cela, la description précise du phénotype reste la donnée primordiale à recueillir, accompagnée d'informations de généalogie. L'une des premières actions menées par l'Observatoire a donc été l'élaboration et la diffusion d'une fiche de déclaration et de description des anomalies constatées dans les élevages. La fiche propose un descriptif assez simple d'éléments morphologiques, répartis par grands compartiments anatomiques, dont des éléments peuvent être anormaux. Cette fiche permet aux différents acteurs, quel que soit leur niveau de connaissances et de compétences (éleveur, inséminateur, technicien d'organisme de sélection ou de contrôle de performances, vétérinaires…) de remplir le document, tout en permettant que deux anomalies identiques déclarées de façon indépendante soient assimilées à la même entité pathologique.

Initialement sur support papier, la fiche peut être téléchargée depuis le site de l'ONAB (par la page http://www.onab.fr/Declarer-une-anomalie), ou remplie directement en ligne.

Les partenaires sont fortement sensibilisés pour que toute déclaration s'accompagne de l'envoi de prélèvements biologiques (sang total, biopsie d'oreilles, voire autre tissu) pour extraction d'ADN, sans lesquels aucune étude génétique n'est possible. Tous les prélèvements sont conservés et stockés, même si aucune recherche n'est entreprise immédiatement sur l'anomalie remontée. La centralisation des fiches de données est effectuée à l'Institut de l'Elevage, l'objectif étant de collecter suffisamment d'éléments remontant du terrain pour que l'ONAB puisse être lanceur d'alerte en cas d'anomalie émergente. La veille scientifique et technique opérée par l'ONAB se traduit également par la communication rapide auprès des acteurs de la génétique de la détection d'une anomalie à l'étranger dans une race utilisée ou sélectionnée en France.

D. Inventaire des anomalies génétiques connues chez le bovin

Les anomalies d'origine génétique dans l'espèce bovine sont étudiées depuis plus d'un siècle, avec un nombre important de cas qui ont fait l'objet de publications dès le début du 20ème siècle. Certains chercheurs se sont très tôt attachés à synthétiser l'information scientifique concernant ces anomalies. Aujourd'hui, une source documentaire importante est la base de données Online Mendelian Inheritance in Animals (OMIA) qui concerne l'ensemble des espèces animales (http://omia.angis.org.au/home/). Au 15/05/2016, on dénombrait 480 entrées pour l'espèce bovine.

Cette base vient compléter celle existant pour l'homme (OMIM, http://omim.org) avec 4757 entrées d'anomalies à déterminisme connu et plus de 3000 autres au déterminisme non encore élucidé, et celle décrivant toutes les ressources de la souris (MGI, http://www.informatics.jax.org/).

Les anomalies sont très diverses et peuvent être classées selon différents critères :

- La fonction atteinte
- Le stade physiologique auquel l'anomalie est exprimée
- Son déterminisme génétique (récessif, dominant, …)
- La nature du polymorphisme responsable (SNP, Indel, variant structural) et l'impact sur le ou les gènes concernés
- Sa distribution dans les races et la fréquence de l'anomalie

Les informations disponibles ont beaucoup évolué ces dernières années, grâce aux avancées de la génétique moléculaire et des techniques d'analyse du génome (Nicholas et Hobbs, 2014). Ainsi, pendant de nombreuses années, les bases ne contenaient que l'inventaire et la description clinique des anomalies ainsi que l'hypothèse génétique émise sur la base de la transmission observée, mais sans base physiologique ou moléculaire associée. Les anomalies complètement caractérisées, comme le BLAD ou le CVM restaient des exceptions. Depuis une dizaine d'années, au contraire, le nombre d'anomalies complètement caractérisées augmente rapidement.

II. Evolution des technologies de génotypage et séquençage : intérêt en génétique des bovins

Ces 25 dernières années, de profonds bouleversements technologiques ont permis une augmentation importante du débit et du volume des informations obtenues, ainsi qu'une baisse toute aussi importante des coûts associés aux outils utilisés en génétique animale et végétale. Dans cette partie nous présenterons l'évolution de ces outils et de leurs caractéristiques.

A. Génotypages :

Dans la plupart des situations, la détermination d'un gène candidat à partir de la seule information clinique des anomalies est très peu fiable. Les généticiens utilisent une étape de cartographie sur le génome sans a priori sur la nature et le mode d'action du gène mis en cause. Pour cela, ils utilisent les principes de la liaison ou de l'association génétique entre l'anomalie et des marqueurs. Les deux approches reposent sur la proximité chromosomique entre la mutation responsable de l'anomalie et les marqueurs, et donc sur leur co-ségrégation. La liaison se traduit par la co-transmission intra-famille chez des produits issus d'un ou de deux parents hétérozygotes. Si un parent est double hétérozygote MA/ma

pour les allèles A et a de l'anomalie et pour les allèles M et m d'un marqueur, ces deux locus étant proches et ayant un taux de recombinaison faible, la plupart des produits présentant l'anomalie a auront reçu l'allèle m, indiquant que l'anomalie se trouve sur le génome à proximité du marqueur. L'association est un principe relativement similaire mais à l'échelle de la population, la fréquence f(am) de l'haplotype am étant très supérieure au produit des fréquences alléliques f(a).f(m). L'association présente plus de risque de faux positif (car les causes d'association sont multiples) mais elle est plus puissante et plus résolutive (car une association à l'échelle de la population suppose une distance plus courte qu'une liaison intra famille).

Aucune hypothèse n'étant faite sur la fonction des marqueurs qui sont supposés neutres, le choix des marqueurs utilisés a été dicté par des arguments génétiques et pratiques : la couverture du génome, l'informativité, la facilité et le coût du génotypage. L'informativité dépend du nombre d'allèles et de leur fréquence. Les variants recherchés pouvant être n'importe où sur le génome, les marqueurs sont choisis pour couvrir tout le génome, avec une densité suffisante. Pour cela, il faut bien sûr connaitre la position des marqueurs sur le génome et leur position relative entre eux. Différentes cartes ont été construites par le passé, par exemple les cartes génétiques (basées sur une distance fonction du taux de recombinaison), les cartes d'irradiation (basées sur la conservation de petits segments d'ADN dans des hybrides irradiés), ou les cartes physiques (basées sur l'alignement de BAC). Les marqueurs sont aujourd'hui positionnés sur la séquence du génome (ie. paragraphe II-B-1), la distance étant donc le nombre de bases. Les marqueurs utilisés ont varié au cours du temps, principalement en fonction des techniques de génotypage, de leur multiplexage et de leur automatisation, et donc de leur coût.

1. Les différents types de marqueurs

Différents types de marqueurs ont été utilisés chez le bovin, on en retiendra trois types principaux : les marqueurs RFLP (*Restriction Fragment Length Polymorphism*), les micosatellites et les SNP.

Les marqueurs RFLP utilisent le polymorphisme de taille d'un fragment d'ADN amplifié par PCR (*Polymerase Chain Reaction*) après digestion par des enzymes de restriction et migration des produits par électrophorèse sur gel. Les différents allèles d'un marqueur sont repérés par la taille des fragments d'ADN obtenus qui dépendent de la présence ou l'absence d'un polymorphisme chez un individu dans le site de restriction de l'enzyme. Les marqueurs microsatellites appartiennent à une sous-catégorie des séquences répétées du génome. Ils sont définis par la répétition successive d'un motif de 2 à 5 nucléotides. Ces marqueurs sont très polymorphes et leurs allèles se distinguent en fonction du nombre de répétitions et donc de la taille des fragments après PCR et migration sur gel de polyacrylamide. Une certaine automatisation est possible grâce au multiplexage PCR (5 à 10 marqueurs amplifiés simultanément) et à la migration sur séquenceur capillaire. Ils ont été largement utilisés entre 1990 et 2005 environ. Enfin, les marqueurs SNP (*Single Nucleotide Polymorphisms*) consistent en une différence d'un nucléotide entre deux individus à une position donnée de leur séquence d'ADN. Par

rapport aux autres types de marqueurs, les SNP sont beaucoup moins polymorphes (en général deux allèles et jusqu'à quatre pour chaque position), ce qui diminue l'information au niveau d'un site, mais beaucoup plus nombreux à l'échelle du génome, chaque base du génome pouvant être modifiée par rapport à une référence fixée. En outre leur génotypage est plus facilement « multiplexable » (cf. ci-dessous).

2. Les puces de génotypage : accès au génotypage haut débit

L'obtention de génotype pour un nombre élevé de marqueurs par individu en un temps raisonnable et à un coût abordable représente une contrainte importante dans le choix d'utilisation des marqueurs moléculaires. Différentes techniques ont été développées, les premières encore très manuelles pour chaque individu, et limitées en capacité de multiplexage des marqueurs et des échantillons (Vignal *et al.*, 2001).

Le développement des puces de génotypage au cours des années 2000 a révolutionné les capacités de débit de typage et engendré une utilisation privilégiée des SNPs en tant que marqueurs moléculaires.

a) Principe de fonctionnement:

La technologie des puces à SNP est fondée sur une adaptation des puces à ADN, dont le principe repose sur la propriété d'hybridation spécifique de la molécule d'ADN et qui permettent la capture spécifique d'une région d'intérêt du génome. Elles ont été développées dans un premier temps pour l'étude de l'expression quantitative des gènes (Schena *et al.*, 1995) avant d'être adaptées à d'autres utilisations (Fan *et al.*, 2003). La technologie de génotypage la plus utilisée chez le bovin actuellement est proposée par Illumina et fait appel à la chimie de type *Infinium* assay (Gunderson *et al.*, 2005 ; Steemers *et al.*, 2006). Une puce est composée d'une lame de silice plane, gravée d'une matrice de micro-puits dans lesquels sont aléatoirement réparties des microbilles de silice (*Beads*). Chaque microbille est le support d'oligonucléotides identiques (*Bead Type)* contenant une séquence de décodage et une amorce de 50 nucléotides synthétisée spécifiquement pour correspondre à la séquence complémentaire ou séquence flanquante au niveau de la position des polymorphismes.

En fonction du type de SNP à génotyper, deux designs différents des amorces sont utilisés :

- le design Infinium I utilise deux amorces par SNP donc deux types de microbilles, chacune des amorces correspondant à un allèle (Gunderson *et al.*, 2005 ; www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_iselect_design.pdf) ;
- le design Infinium II utilise une seule amorce par SNP. Cette amorce s'arrête à la base précédant le SNP et permet le génotypage des deux allèles sur un seul type de microbilles (Steemers *et al.*, 2006 ; www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_iselect_design.pdf).

Selon le protocole Illumina (*Figure 1*), le génotypage d'un échantillon commence par une étape d'amplification PCR non-spécifique de l'ADN génomique de l'individu. Les produits d'amplification subissent ensuite un fractionnement enzymatique, avant d'être concentrés et chargés sur la puce. La détermination spécifique des allèles pour un SNP s'effectue en deux étapes : i) l'hybridation spécifique des molécules d'ADN amplifiées avec les amorces ; ii) l'extension enzymatique avec une base marquée (ie. *Allele specific primer extension* pour *Infimium I* et *Single base extension* pour *Infinium II*). Deux couleurs de marquage fluorescent, vert et rouge, sont utilisées pour différencier les bases intégrées. Le génotype est déterminé par les proportions des différentes intensités observées : une majorité de signaux vert ou rouge indique un génotype homozygote, tandis qu'une proportion équivalente de signaux vert et rouge (couleur jaune) indique un génotype hétérozygote. Pour obtenir les résultats, la puce est scannée par un système d'imagerie automatique à haute résolution qui enregistre pour chaque position de la matrice, c'est-à-dire chaque microbille, les signaux de fluorescence émis et leur intensité. De façon à disposer de résultats fiables, le test de chaque SNP est répliqué sur une vingtaine de billes en moyenne. Les billes avec leurs oligonucléotides étant étalées aléatoirement sur les lames, la position de chaque SNP, qui est spécifique de chaque lame, est identifiée par le fabricant par une série d'hybridations successives, en utilisant non pas la séquence flanquante du SNP mais la partie réservée au décodage. Les signaux d'hybridation captés par le scanner sont ensuite analysés par le logiciel *Genome Studio®*, développé par Illumina. Les résultats de chaque bille sont consolidés par SNP puis sont interprétés par recherche de clusters. Idéalement, les trois génotypes se répartissent en trois clusters disjoints (*cf.* Figure1*, étape 5*). Les points en dehors des clusters sont interprétés comme des typages indéterminés et donc manquants. Lorsque le nuage de points (exprimés en coordonnées polaires pour la présentation graphique) ne permet pas de distinguer des clusters suffisamment disjoints, le marqueur n'est pas interprétable.

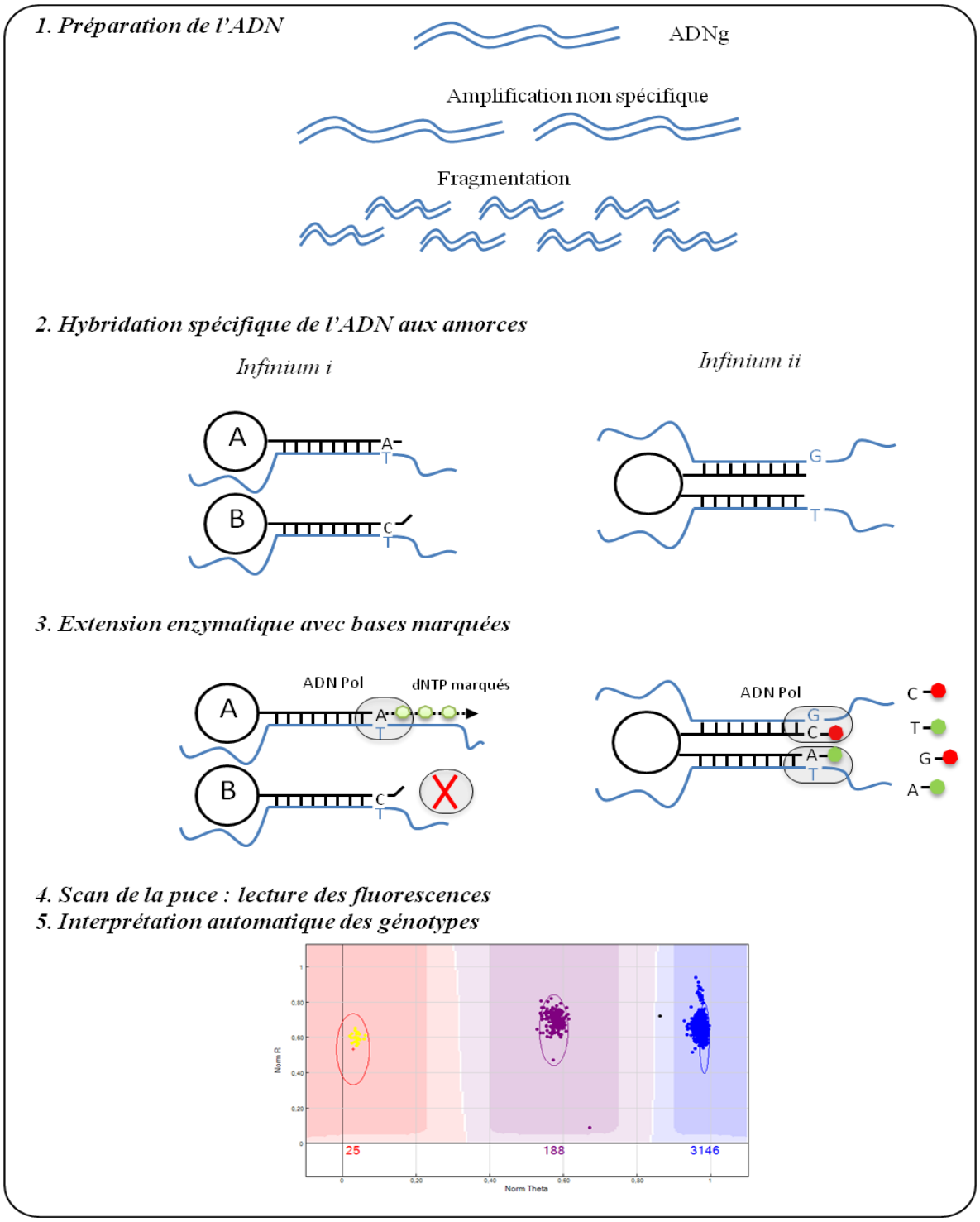

Figure 1: Procédé général du génotypage d'un marqueur SNP sur une puce Illumina utilisant la technologie Infinium assay.

L'ADN de l'individu est amplifié puis fragmenté et chargé sur la puce (étape 1). Les fragments s'apparient spécifiquement à l'amorce correspondante (étape 2). Une fois appariés, il y a une réaction d'extension d'une base avec des nucléotides marqués (étape 3). Les bases insérées sont révélées par activation du fluorochrome. Le signal résultant est lu et analysé par le logiciel Genome Studio, qui renvoie automatiquement le génotype du marqueur SNP pour chaque individu (étapes 4 et 5). La figure est adaptée de Gunderson et al.(2005) et Steemers et al (2006).

b) Avantage des puces à SNP:

Les avantages de la technologie de génotypage par puce sont majeurs :

- automatisation totale des analyses ;
- typage de plusieurs milliers et même de centaines de milliers de marqueurs en parallèle ;
- taux de résultat (« call rate ») élevé (>99% dans une forte proportion de cas) ;
- taux d'erreur très faible, de l'ordre de 0.02% pour des marqueurs connus et optimisés ;
- multiplexage d'échantillons, par exemple par 8, 12, 24 ou 96 selon la densité de marqueurs ;
- au final, un coût par marqueur très faible et un coût par échantillon (le critère essentiel) relativement abordable.

Cette automatisation, l'importance des investissements, et la nécessité de réaliser beaucoup d'analyses pour réduire les coûts ont également été à l'origine de la création de plateformes spécialisées, de sorte que le travail de génotypage est de moins en moins réalisé par les équipes de recherche et généralement externalisé. Ainsi, tous nos génotypages ont été réalisés par Labogena.

Le développement de ces nouvelles technologies débloque le verrou lié au génotypage des marqueurs. Il permet une automatisation complète du génotypage des SNP, apportant un avantage par rapport aux marqueurs microsatellites. En quelques jours ou semaines, on accède aux typages, c'est-à-dire aux allèles portés par chaque individu analysé, pour des milliers de SNP en même temps pour un coût réduit.

c) Types de puces proposées chez le bovin:

Chez le bovin, après plusieurs puces expérimentales développées à partir de 2005 par Affymetrix et Illumina, la première puce SNP libre de droit, accessible en terme de prix et très satisfaisante techniquement a été développée par Illumina et mise sur le marché à partir de 2007. Depuis, différentes versions et types de puces à SNP ont été déclinés et utilisés jusqu'à aujourd'hui:

- **La puce Illumina BovineSNP50** dite de moyenne densité, avec deux versions successives proches contenant environ 54 000 marqueurs SNP. Sa densité médiane est d'un marqueur tous les 37,4 kb, avec un minimum de 20kb et rarement plus de 100kb entre deux marqueurs (Matukumalli *et al.*, 2009). Elle permet de typer 24 animaux simultanément.
- **La puce haute densité** (dite HD) contenant 777 000 SNP avec une densité moyenne d'un marqueur tous les 4.5 kb, permettant de typer 8 animaux simultanément. Une puce Affymetrix Axiom a également été conçue en 2011 avec environ 600 000 marqueurs.
- **La puce basse densité** (LD) contient environ 7000 marqueurs choisis de façon à optimiser l'imputation et le phasage des génotypes sur les marqueurs de la puce 50K dans le cadre du développement de la sélection génomique (Boichard et *al.*, 2012a). Cette puce a été la première à permettre l'ajout de SNP avec design particulier, en plus du panel de base. Elle a été la base par exemple de la puce du consortium EuroGenomics.

B. Séquençage de génomes

Comme pour le génotypage, différentes technologies ont été successivement développées dans le but de connaitre l'enchaînement des bases constituant la séquence des gènes et du génome.

1. Séquençage de première génération : le séquençage Sanger

a) Principe

Dans la méthode Sanger, une molécule d'ADN subit une réaction d'élongation en utilisant l'ADN à séquencer comme matrice. Les bases fournies pour cette réaction sont soit des bases normales et non marquées, soit des bases modifiées (di-desoxyribose) et marquées. Le processus d'élongation se poursuit tant que des bases normales sont incorporées et il est stoppé lors de l'incorporation d'une base modifiée. Les proportions des bases normales et modifiées sont optimisées de sorte que tous les fragments coexistent dans le milieu de réaction. Les fragments sont ensuite séparés par électrophorèse sur la base de leur longueur. Avec 4 couleurs de marquage, une par base, on identifie ainsi la dernière base incorporée de chaque fragment et donc la série de bases constituant la séquence. Cette méthode garde une précision acceptable jusqu'à 400 à 600 bases. C'est encore aujourd'hui la méthode de référence pour les petits débits de séquençage.

b) Applications majeures : séquençage et alignement du génome de référence

L'amélioration des technologies de séquençage et le développement des outils de bioinformatique nécessaires à l'assemblage *de novo* de génomes ont ouvert la voie à l'obtention des séquences de génome complexes, en particulier des mammifères. Ainsi, après la séquence du génome humain, les années 2000 ont été marquées par la publication successive de la séquence des génomes de référence de plusieurs espèces domestiques dont le bovin. Les données de séquence ont été produites par le *Baylor College of Medicine* à partir du séquençage Sanger de l'ADN génomique de la femelle Dominette de race Hereford selon une stratégie *Shotgun* (découpage de l'ADN en fragments aléatoires, amplifiés par PCR puis séquencés) et de banques de BAC (*Bacterial Artificial Chromosomes*) constituées à partir de l'ADN de son père (Domino). Ces travaux ont conduit à la l'obtention du premier assemblage de génome de référence complet disponible en 2006 et publié en 2009 (Elsik *et al.,* 2009).

Différentes versions de l'assemblage du génome ont depuis été publiées en fonction des processus d'amélioration de la qualité de l'assemblage réalisé (correction d'erreurs d'assemblage, augmentation de la couverture de séquençage …). Le génome de référence bovin actuellement utilisé, UMD 3.1, est la troisième version de l'assemblage publié par Zimin *et al.,* (2009). Il correspond à une taille de 2,65 Gb et l'alignement correct d'environ 91% du génome bovin. Une version plus récente de l'assemblage du génome bovin a été publiée fin 2015 (Btau5). Cette dernière a bénéficié des améliorations des données apportées par les technologies de séquençage de troisième génération qui augmentent la qualité

d'assemblage grâce à l'obtention de lectures longues, augmentant la qualité de l'alignement au sein des séquences répétées. Malgré sa qualité, elle est encore peu utilisée puisque le passage d'un assemblage à un autre nécessite d'importants efforts tels que le réalignement des séquences de tous génomes individuels déjà séquencés, la reprise à zéro des cartographies génétiques, …

L'assemblage du génome bovin et les efforts de séquençage qui l'ont accompagné ont offert aux généticiens l'accès à de nombreux marqueurs moléculaires et à leur position physique sur le génome (carte physique). Ainsi, dès 2006, la comparaison des séquences partielles disponibles pour d'autres animaux (séquences d'ADNc, shotgun à faible couverture, séquences de BAC, etc) avec le génome de référence a permis de mettre en évidence près de 3,2 millions de SNP, disponibles dans la base publique dbSNP. Ces variants ont constitué la matière première utilisée pour la conception des puces SNP présentées précédemment.

2. Séquenceurs de seconde génération : le séquençage haut débit

L'obtention de la séquence du génome complet d'autres individus est peu abordable par l'intermédiaire du séquençage Sanger qui demande une grande quantité d'ADN pour obtenir une couverture suffisante et surtout qui est très coûteuse pour de grandes séquences. L'arrivée des séquenceurs de seconde génération a permis de diminuer les coûts de séquençage d'une base et le développement du séquençage partiels ou total de génome.

a) Principe de séquençage de seconde génération

Dans la technologie Illumina utilisée dans les séquenceurs Genome Analyzer puis HiSeq, le principe du séquençage Sanger est conservé mais il est miniaturisé et parallélisé. Un grand nombre de segments à séquencer, de quelques centaines de bases, sont fixés par les deux extrémités sur une lame, multipliés en « buisson » pour disposer d'un signal suffisant et subissent une réaction d'élongation comme en Sanger. A chaque élongation d'une base, un signal lumineux est émis et lu par imagerie. Puis, la base est restaurée afin de permettre la poursuite d'un nouveau cycle d'élongation du brin d'ADN. La réaction d'élongation se produit successivement par les deux extrémités des fragments d'ADN, ce qui permet des lectures appariées (ou pair ends). Au cours des évolutions de technologies, la longueur des lectures obtenues a augmenté d'une trentaine de bases dans les premières machines à 125-150 bases dans les séquenceurs HiSeq et même plus de 400 dans les MySeq. Par ailleurs, la densité de « *buissons* » a augmenté pour atteindre plus de 150 millions par site, permettant aujourd'hui d'obtenir à chaque « run » près de 50 gigabases de séquences.

b) Traitement des données NGS : alignement des données de séquence et détection des variants

Les NGS produisent un très grand nombre de séquences courtes. Ces données posent de gros problèmes méthodologiques pour réaliser un assemblage d'un génome complet *de novo*, même si la combinaison

d'algorithmes très performants et de librairies de segments de longueur variée a récemment permis des avancées spectaculaires (génome du blé, par exemple).

En revanche, avec l'existence d'une séquence de référence, cette technologie est devenue très attractive pour le re-séquençage de génome de nouveaux individus. En effet, il est facile de reconstituer le génome d'un individu en réalisant un alignement des lectures obtenues par rapport au génome de référence de l'espèce et en identifiant les variants portés par comparaison. L'interprétation des données de séquence a nécessité le développement d'outils bio-informatiques pour le traitement de l'ensemble de ces données. Par exemple l'alignement des lectures sur le génome de référence est réalisé avec le logiciel BWA (Li et Durbin, 2009) tandis que les outils SamTools (Li *et al.*, 2009) ou GATK (McKenna *et al.*, 2010) permettent l'identification des petites variations de type SNP ou les insertions ou délétions de quelques paires de bases.

Les variations observées sont de nature très diverse. Les plus simples à observer sont les SNP. Dans ce cas, on observe des lectures alignées au même endroit et différant d'une seule base. Si la couverture est suffisante, on déduit qu'un individu est homozygote référence si ses lectures sont identiques à la référence, homozygote alternatif si ses lectures sont toutes différentes de la référence et hétérozygote si deux types de lecture sont observés. Les faux positifs peuvent avoir diverses causes ; par exemple les erreurs de séquençage et les erreurs d'alignement, en particulier en cas de séquence dupliquée. Les petites insertions et délétions nécessitent un travail plus compliqué car elles sont à l'origine de décalages entre séquences, ce qui nécessite des réalignements locaux pour éviter de conclure à de nombreuses différences simplement liées à ces décalages. Enfin, les variants structuraux sont plus complexes à mettre en évidence et nécessitent des approches spécifiques. On peut citer, par exemple, du moins précis au plus précis : la variation de couverture qui laisse supposer des délétions ou des duplications ; la cartographie à deux endroits éloignés des deux lectures de la paire, attendues proches ; une lecture n'alignant nulle part sur la référence mais dont deux fragments s'alignent sur des régions distinctes sur la référence du génome. Cette dernière méthode dit « *split-read* » est la plus précise car elle permet de déterminer le point exact de cassure.

c) Prédiction de l'effet des variations observées

Nos connaissances sur l'impact potentiel d'une variation sont encore fragmentaires. Une première classification peut être réalisée en fonction de la position sur le génome et par rapport aux gènes. On distingue ainsi les variants intergéniques (pour lesquels l'annotation est souvent pauvre ou absente), des variants intragéniques, eux-mêmes répartis en variants dans les exons, introns, régions régulatrices en amont ou en aval. Lorsqu'ils affectent la séquence protéique, leur annotation est souvent précise. Pour ce qui concerne les anomalies génétiques, on se concentre souvent sur les variants qui affectent la région codante d'un gène et entraînent une modification de la protéine, car ils ont souvent des conséquences biologiques plus graves que celles qui se situent en dehors du codant. Cette affirmation compte beaucoup d'exceptions mais elle est suffisamment vérifiée pour que les variants dans la partie codante du génome

soient étudiés en priorité. Ainsi, on cible tout particulièrement les mutations (i) ajoutant ou supprimant des codons stop, (ii) entraînant un décalage du cadre de lecture, (ii) modifiant le site d'initiation ou un site d'épissage, (iv) induisant une substitution d'acides aminés avec des propriétés très différentes, ou bien encore (v) affectant une région très conservée de la protéine entre espèces. Les logiciels de prédiction des fonctions des protéines comme SIFT (Kumar *et al.,* 2009) ou Polyphen (Adzhubei *et al.,* 2010) permettent d'orienter les choix avec efficacité, tandis que Ve!P (Variant Effect Predictor) est un outil intégré d'annotation (McLaren *et al.,* 2016)

Pour conclure, les NGS ont modifié considérablement les méthodes d'étude du génome en général et des anomalies génétiques en particulier. Elles fournissent une très grande quantité de données, sur l'ensemble du génome et permettent de déduire le génotype d'un individu pour l'ensemble des variants de son génome. Par ailleurs, le coût a beaucoup diminué. S'il reste encore sensiblement plus élevé (environ 10 à 30 fois plus) que le génotypage, il devient suffisamment abordable pour que le séquençage d'un individu soit l'option de loin la plus rapide et la moins chère pour rechercher un variant particulier dans son génome.

III. Identification des anomalies génétiques : du phénotype au gène et du gène au phénotype

A. Les ressources d'information disponibles chez le bovin et leur utilisation dans les travaux d'identification des anomalies génétiques

L'équipe G2B dans laquelle j'ai effectué ma thèse, et plus précisément l'une de ses composantes, l'UMT 3G, réalise les évaluations génétiques et génomiques pour l'ensemble des races françaises. A ce titre, elle a accès à un certain nombre de ressources d'intérêt pour l'étude des anomalies génétiques, auxquelles viennent s'ajouter d'autres données générées dans le cadre de projets de recherches. La nature de ces ressources et l'utilisation qui peut en être faite seront détaillées dans cette partie.

1. Pedigrees et état civil bovin

L'état civil bovin comprend l'identifiant national unique, la race d'appartenance, les dates de naissance et de mort, ainsi que les informations de père et mère (lorsqu'ils sont renseignés) de tous les animaux nés en France depuis les années 1950 environ. Ces informations permettent notamment la reconstruction des haplotypes à partir des données de génotypage, l'identification des ancêtres fondateurs des races, les calculs de probabilité de transmissions d'allèles, la recherche d'ancêtres communs sur les voies paternelle et maternelle des animaux d'anomalies récessives,…

2. Les phénotypes

Des données phénotypiques pour une quarantaine de caractères de production, conformation, reproduction et santé sont mises à disposition dans chacune des races pour la réalisation des évaluations génétiques ou génomiques. Ces phénotypes sont issus des contrôles de performances en ferme (contrôle laitier, bovin croissance), des pointages morphologiques réalisés par les Organismes de Sélection, ou sont renseignés par les éleveurs et les techniciens d'insémination. Leur utilisation dans le cadre d'une approche de génétique inverse (cf infra) permet de tester l'effet des mutations étudiées sur une grande catégorie de fonctions biologiques et ainsi d'améliorer notre connaissance de la fonction de certains gènes encore peu étudiés.

3. Les données génomiques

Les données de génotypage de plus de 500 000 individus (principalement de races Holstein, Montbéliarde, Normande, Brune, Charolaise, Limousine et Blonde d'Aquitaine) nettoyées, phasées et imputées pour les marqueurs de la puce Illumina BovineSNP50 dans le cadre de la sélection génomique française sont disponibles pour la réalisation de recherches. Cette ressource permet de disposer des génotypes d'animaux reliés aux individus atteints (leurs pères par exemple), d'avoir accès à des données de génotypage sur une large population pour des mutations d'intérêt, de disposer d'une population contrôle, d'effectuer des recherches de déficit en homozygotes (cf infra), de bénéficier du traitement automatique des données, etc.

D'autre part, dans le cadre de différents projets de recherche notre équipe a reséquencé les génomes de plus de 350 animaux d'une dizaine de races (Boussaha *et al.,* 2016). Elle a par ailleurs accès aux génomes de plus de 1000 individus à travers sa participation au consortium 1000 génomes bovins (Daetwyler *et al.*, 2014). Ces données constituent une ressource de choix pour la recherche de mutations délétères dans les génomes des principaux fondateurs des races et la réalisation d'études de génétique inverse (cf infra).

4. Les annotations du génome

Différentes bases de données d'annotation fonctionnelle du génome bovin sont accessibles par l'intermédiaire de « *genome browsers* ». Les plus connus sont l'UCSC (http://genome-euro.ucsc.edu/), le NCBI (https://www.ncbi.nlm.nih.gov/) et Ensembl (http://www.ensembl.org/). Ils donnent un accès facilité à la localisation des gènes, la position des exons et introns, l'existence des transcrits, de variations connues et permettent aussi une comparaison avec les gènes orthologues connus chez d'autres espèces. Ces *genome browsers* se basent sur deux sets d'annotations du génome bovin : la base d'annotation Nord-américaine RefSeq développée par le NCBI (O'Leary *et al.*, 2016), et la base européenne Ensembl (Bronwen *et al.*, , 2016). Chaque set de transcrits est construit à partir de pipeline

d'annotation du génome qui positionne les gènes en utilisant les données de séquençage d'ADNc ou d'EST pour les gènes dont les transcrits ont été étudiés, ou bien en réalisant une prédiction de la position du gène par comparaison de la séquence génomique avec les séquences de gènes connus dans d'autres espèces.

B. Les méthodes de détection des anomalies récessives

Les approches classiques de génétique s'intéressent dans un premier temps aux phénotypes pour ensuite rechercher les mutations responsables dans le génome des animaux présentant ces mêmes phénotypes. Nous présenterons dans un premier temps le principe de cartographie par homozygotie, méthode la plus souvent utilisée pour les anomalies récessives. Avec le développement des typages et de la sélection génomique, les travaux sur haplotypes et notamment la recherche de déficit en homozygotes se sont également développées pour mettre en évidence des anomalies responsables de mort embryonnaire ou juvénile non détectée par les observatoires. Enfin, si l'on considère les bases de données de séquence, d'autre méthodes sont actuellement possibles pour repérer les anomalies directement dans les génomes des individus fondateurs, et ce avant qu'elles ne soient rapportées sur le terrain. Ce sont les approches de screening des données de séquence et de détection des anomalies par une méthode de génétique inverse qui ont également fait l'objet de deux études dans cette thèse et dont nous présenterons en dernier lieu le principe.

1. Du phénotype au gène : la cartographie par homozygotie

La cartographie par homozygotie, proposée par Lander et Bostein (1987) et revisitée par Charlier *et al.* (2008) est la stratégie la plus utilisée pour le clonage positionnel de mutations récessives. Son principe repose sur le suivi de la transmission du fragment de chromosome entourant la mutation entre individus apparentés en s'aidant des marqueurs moléculaires. En effet, lorsqu'une mutation apparait chez un individu, elle est transmise à sa descendance avec un fragment du chromosome qui l'entoure. Au fil des générations, ce fragment de chromosome, transmis à l'identique, se réduit peu à peu en fonction des évènements de recombinaison. Les individus atteints, homozygotes à la mutation, ont reçu par chacun de leurs parents le segment associé à la mutation qui provient du même ancêtre commun, chez qui est apparue la mutation (supposée unique). Le suivi de ces fragments chromosomiques transmis avec une mutation délétère est assuré par l'intermédiaire des marqueurs. Par comparaison des génotypes des individus atteints, la position de la mutation responsable de l'anomalie est repérée par l'identification du plus grand segment homozygote commun à l'ensemble des individus atteints et jamais (ou rarement) homozygote chez des individus sains (*Figure 2*).

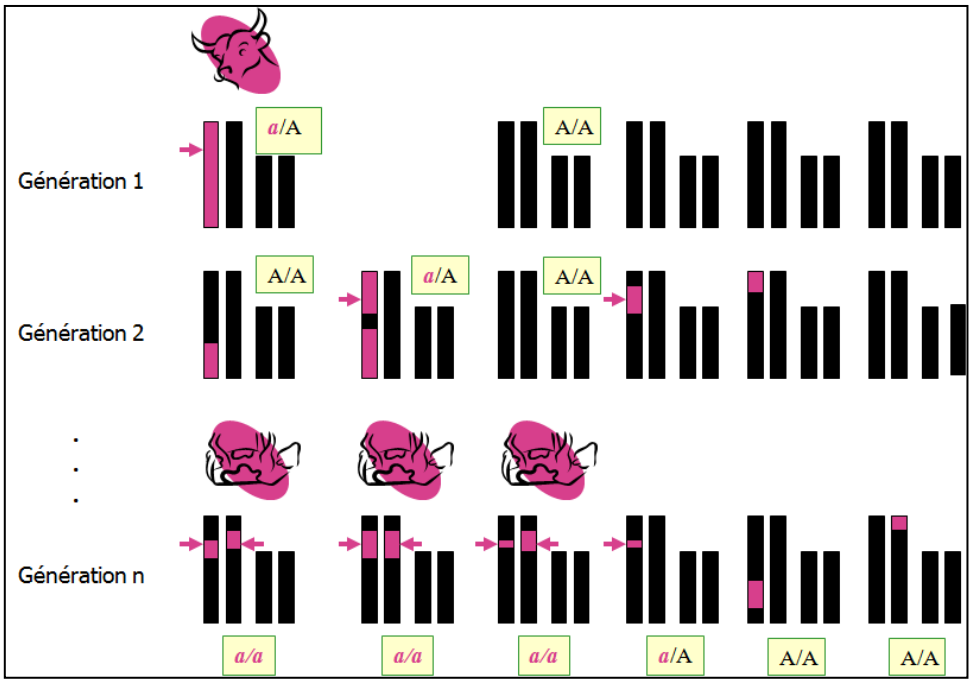

Figure 2: Illustration de la cartographie par homozygotie.

Une mutation récessive apparait chez l'ancêtre (emplacement de la flèche sur le chromosome rouge). Cette mutation se transmet à l'état hétérozygote à certains de ses descendants phénotypiquement normaux. Après quelques générations, les premiers individus consanguins apparaissent, dont certains, homozygotes à la mutation (les trois individus colorés à la génération n sur la figure), sont atteints et révèlent l'émergence. Autour de la mutation, on observe un segment chromosomique conservé provenant de l'ancêtre ayant transmis l'anomalie, que l'on utilise pour cartographier la mutation.

Lander et Bostein (1987) indiquaient que la méthode serait efficace à partir d'un faible nombre d'individus, grâce à l'utilisation de cartes génétiques saturées en marqueurs moléculaires. Chez le bovin, la cartographie par homozygotie s'est principalement développée grâce aux puces à SNP. L'augmentation forte de la densité de marqueurs le long du génome, en passant d'une densité de couverture d'environ 4000 marqueurs microsatellites (Iraha *et al.*, 2004) (dont seulement quelques centaines génotypés en routine) à environ 54 000 marqueurs SNP (soit un marqueur tous les 37,5 kb en moyenne) avec le développement des puces (cf. paragraphe précédent), permet le suivi précis des haplotypes identiques par ascendance (IBD) chez tous les individus atteints de l'anomalie. Charlier *et al* (2008) démontrent l'efficacité de cette méthode de cartographie chez le bovin avec localisation de 5 anomalies génétiques à partir des données de génotypage de trois à douze individus atteints.

Une fois la région localisée, il faut identifier le gène et la mutation responsables. La recherche d'un gène candidat s'appuie depuis longtemps sur les possibilités de comparaison à partir des bases de données de connaissances fonctionnelles et de phénotypes pour d'autres espèces comme l'homme ou la souris par exemple (cf. bases OMIM, MGI et OMIA, présentées précédemment cf. III-A-5).

Avec les NGS et l'accès au génome des animaux atteints, cette étape a évolué vers de nouvelles stratégies par screening du génome et un filtre des variations identifiées dans l'intervalle par rapport aux données de séquence d'animaux contrôle sains. Enfin, parmi les variants candidats positionnels, la sélection finale repose sur l'annotation fonctionnelle. Le variant le plus probable fait ensuite l'objet d'une validation statistique par génotypage à grande échelle et de tests fonctionnels (analyse d'ARN, de protéines, immunohistochimie, voire dans certains cas création de modèles transgéniques). Ces stratégies ont été employées pour l'étude de deux anomalies récessives dont les travaux sont présentés dans les chapitres 2 et 3 de ce document.

2. Détection de déficits en homozygotes à partir de données de génotypage

Le principe de cartographie par homozygotie peut être adapté pour cartographier des mutations récessives délétères à partir des bases de données de génotypage de sélection génomique et sans données phénotypiques. Toutefois, l'approche présente deux grosses différences : d'une part, elle nécessite des effectifs d'animaux génotypés très élevés ; d'autre part, les marqueurs isolés étant peu informatifs, il faut disposer d'haplotypes en tout point du génome et la procédure nécessite donc une étape préalable de phasage.

Le principe est le suivant : si un haplotype ségrège à une fréquence relativement élevée dans la population, mais n'est pas observé à l'état homozygote dans cette même population, on en déduit que les animaux homozygotes à l'haplotype correspondant n'ont jamais été candidats à la sélection. Soit parce que ces animaux n'existent pas en raison d'une mortalité au cours de la gestation (avortement précoce ou tardif) ou bien néo-ou péri natale, soit parce qu'ils présentent des caractéristiques éliminatoires pour la sélection. L'haplotype ainsi repéré est supposé être associé à une mutation délétère récessive. Le scan des haplotypes observés dans les typages des animaux candidats à la sélection permettent de repérer ces déficits significatifs en homozygotes.

Cette approche, initialement proposée par VanRaden *et al.,* (2011) a été reprise dans plusieurs études ces dernières années en particulier dans les races laitières où la quantité de données de génotypage permet la détection efficace de tels haplotypes, même avec une fréquence faible (fréquence entre 1 et 5%*)* (Fritz *et al.*, 2013 ; Sahana *et al.*, 2013 ; Sonstegard *et al.,* 2013 ; Daetwyler et al., 2014 ; McClure *et al.*, 2014 ; Venhoranta *et al.*, 2014 ; Pausch *et al.,* 2015 ; Adams *et al.*, 2016 ; Menzi *et al.,* 2016 ; Schütz *et al.*, 2016 ; Schwarzenbacher *et al.*, 2016). Une fois ces régions identifiées, l'approche classique consiste alors à prédire le statut (porteur/non porteur) des taureaux d'insémination pour (i) tester l'effet de ces haplotypes sur le taux de conception dans des accouplements à risque (porteur x fille de porteur) versus le reste de la population et (ii) rechercher des mutations candidates dans les données de séquence des principaux contributeurs de chaque race. Au final, ces études ont révélé la présence de mutations récessives et létales à l'état embryonnaire dans chacune des races étudiées et permis d'identifier précisément plus d'une dizaine de mutations de ce genre.

3. Exploitation des données de séquence : le développement des approches de génétique inverse chez le bovin

Enfin, l'accès aux données de re-séquençage d'un grand nombre de génomes grâce à l'essor des NGS permet d'envisager l'étude des mutations responsables d'anomalies génétiques sous un autre angle. Ainsi, ces dernières années ont vu le développement des approches dites de « génétique inverse », dont le principe est d'identifier les variants délétères pour le fonctionnement de la protéine (ie. gain d'un codon stop, décalage du cadre de lecture, modification des sites d'épissage, substitutions d'acide aminé avec un impact fonctionnel…) susceptibles de présenter un fort potentiel pathologique.

Dans ce type de stratégie, l'annotation fonctionnelle de l'effet des variants, les données phénotypiques associées aux gènes chez l'homme, les espèces modèles et domestiques ainsi que les annotations fonctionnelles des gènes sont des ressources essentielles pour établir une prédiction des phénotypes attendus. Après avoir identifié ces variants potentiellement délétères, on cherche ensuite à mettre en évidence ceux qui possèdent un réel effet négatif sur le phénotype des individus.

La confirmation de l'effet des variants sur le phénotype nécessite de repérer des individus homozygotes à l'allèle alternatif délétère. Ces derniers sont le plus souvent identifiés parmi les animaux génotypés dans le cadre de la sélection génomique. Dans les cas où leur fréquence est extrêmement faible, où lorsque leur phénotype est tel qu'ils ne sont pas considérés comme candidats à la sélection génomique, une alternative est d'étudier les produits issus d'accouplement à risque entre individus porteurs. Ce type d'approche a été développé en premier lieux chez l'homme pour l'étude des mutations perte de fonction (on peut citer par exemple les travaux McArthur *et al.,* 2012 ou encore ceux de Sulem *et al.,* 2015). La structure particulière des populations bovines et les programmes de re-séquençage de génomes des taureaux fondateurs de races, rendent possible ces approches qui seront développées dans les chapitres 5 et 6 de cette thèse.

En conclusion, les populations bovines ont un effectif génétique réduit ce qui augmente le risque d'émergence d'anomalies récessives. Des dispositifs de surveillance efficaces ont été mis en place pour repérer rapidement ces nouvelles émergences. Le développement de nouvelles technologies de génotypage et de séquençage associées à la connaissance des génomes bovins et la comparaison avec d'autres espèces donnent des moyens efficaces de résolutions des anomalies émergentes, comme l'illustre le nombre en forte augmentation des mutations causales identifiées ces dernières années. L'étude d'anomalies génétiques à partir de l'exploitation des données de génomes a fait l'objet des travaux réalisés au cours de mes trois ans de thèse. Ces travaux utilisent l'ensemble des technologies modernes de génotypage et de séquençage et illustrent les atouts et inconvénients des trois stratégies présentées précédemment pour la détection d'anomalies récessives chez le bovin.

CHAPITRE 2. APPROCHE USUELLE DE L'ETUDE DES ANOMALIES GENETIQUES AU TRAVERS D'UN EXEMPLE : LE SYNDROME DE CRISE D'EPILEPSIE EN RACE PARTHENAISE

Habituellement, la caractérisation d'une anomalie génétique repose sur trois phases : i) l'identification de l'émergence et la caractérisation clinique du phénotype anormal, ii) la localisation de la région génomique impliquée, et iii) la mise en évidence d'un gène candidat et de la variation délétère responsable. Les outils et stratégies mis en place ont évolué ces dernières années de manière à permettre une gestion rapide et efficace des nouvelles anomalies génétiques émergentes dans les populations. L'identification des émergences et le regroupement des cas se sont progressivement organisés au sein d'observatoires nationaux faisant l'interface entre le terrain et la recherche (Ducos *et al.*, 2003).

La localisation sur le génome a, quant à elle, grandement bénéficié du développement des marqueurs moléculaires et, en particulier à la fin des années 2000, des puces de génotypage à SNP qui permettent de tester rapidement et en une seule étape l'association entre un phénotype et les allèles de milliers de marqueurs répartis sur le génome. En conséquence, la puissance de détection de la mutation responsable d'une anomalie en a été augmentée, comme par exemple dans les approches de cartographie par homozygotie d'une anomalie récessive qui permettent généralement la localisation dans quelques méga-bases sur un chromosome à partir de quelques cas (Charlier *et al.*, 2008).

Le développement des techniques de séquençage a contribué à accélérer les recherches de mutations candidates. Précédemment, on recherchait un gène candidat positionnel, essentiellement sur la base de la génomique comparée en exploitant les bases de données répertoriant les mutations identifiées dans les gènes murins et humains et les phénotypes associés. Puis on séquençait un par un les exons, faute de pouvoir étudier l'ensemble de la région ou même du gène. Aujourd'hui, la stratégie repose essentiellement sur le séquençage du génome de quelques animaux atteints pour identifier des variants candidats par comparaison avec les données de séquence d'individus sains (Daetwyler *et al.*, 2014 ; Murgiano *et al.*, 2014) apportant un gain de temps non négligeable.

Pour illustrer cette situation, ce premier chapitre sera consacré à la description des travaux portant sur un syndrome épileptique d'émergence récente en race Parthenaise. Cette étude constitue un exemple de la mise en place d'une stratégie usuelle de caractérisation d'une anomalie génétique récessive simple, depuis son émergence à l'identification d'une mutation délétère candidate. Elle permet également de montrer les divers avantages des dernières évolutions technologiques avec l'utilisation des données de séquence de génomes complets pour la détection de la mutation candidate.

I. Une mutation dans le gène *MTCL1* est responsable d'un syndrome d'épilepsie récessive en race Parthenaise.

A. Quelques éléments d'introduction sur l'épilepsie

L'épilepsie est l'une des familles d'affections neurologiques les plus fréquentes chez l'homme avec 4 à 10 individus atteints pour 1000 à l'échelle mondiale (OMS). Elle se définit chez un individu comme une prédisposition à générer des crises dites épileptiques répétées dans le temps (Fisher *et al.*, 2014). Ces crises épileptiques résultent d'une activité électrique anormale excessive et hyper-synchrone d'un groupe de neurones ou d'une zone du cerveau et se traduisent par diverses manifestations psychiques et physiques plus ou moins sévères en fonction de la zone cérébrale atteinte (troubles comportementaux, absences, hallucinations sensorielles, convulsions, perte de conscience).

Les causes sous–jacentes aux différentes formes d'épilepsie sont multiples et complexes, mais l'on peut distinguer deux grands types: les épilepsies idiopathiques (ou génétiques) et les épilepsies structurelles ou métaboliques. Les épilepsies dites structurelles proviennent d'une altération de la structure physique des neurones et du cerveau, ce qui induit un fonctionnement électrique anormal. Ces altérations peuvent provenir de lésions acquises (traumatisme, accident vasculaire, infection ou inflammation des tissus, tumeur, carences nutritionnelles, intoxication), ou congénitales et, dans ce cas, liées à un défaut du développement cérébral, lui-même régulé par des facteurs génétiques et environnementaux complexes (Buisson, 2013). Cette catégorie comprend aussi les épilepsies syndromiques, associées à un syndrome à déterminisme génétique simple ou complexe, comme par exemple dans l'épilepsie progressive myoclonique de Lafora (OMIM 254780), les scléroses cérébelleuses, les défauts de métabolisme (Rahman *et al.*, 2013) et dans de nombreux syndromes liés à une anomalie chromosomique (voir Sorge et Sorge (2010) pour revue).

Dans le cas des épilepsies idiopathiques, qui représentent entre 15 et 20% des épilepsies, les crises n'ont pas de sources physiques, structurelles ou métaboliques apparentes. Elles sont souvent la manifestation principale d'une mutation connue ou présumée d'un gène (Jallon et Latour, 2005 ; Berg *et al.*, 2010 ; Shorvon, 2011 ; Berendt *et al.*, 2015). Les familles de gènes impliquées dans les processus de neurotransmission et l'activité électrique neuronale constituent des cibles majeures d'étude, en particulier les canaux ioniques sodium, calcium et potassium ainsi que les neurotransmetteurs et leurs récepteurs (Georges *et al.*, 2004 ; Macdonald *et al.*, 2010 ; Hirose, 2014 ; Helbig, 2015).

En médicine vétérinaire, l'épilepsie constitue également l'une des affections neurologiques les plus courantes chez les petits animaux domestiques. Chez le chien, l'incidence globale est comparable à l'homme et parfois beaucoup plus élevée dans certaines races, ce qui traduit un facteur génétique fort (Berendt *et al.*, 2015). Le gène *ADAM23 (ADAM Metallopeptidase Domain 23)* a été identifié comme

facteur de risque dans les épilepsies idiopathiques, le gène *LGI2 (leucine-rich repeat LGI family, member 2)* est causal dans l'épilepsie juvénile familiale et le gène *NHLRC1 (NHL Repeat Containing E3 Ubiquitin Protein Ligase 1*) est responsable, comme chez l'homme, du syndrome d'épilepsie myoclonique progressive de LaFora (Lohi *et al.*, 2005 ; Seppala *et al.*, 2011 ; Koskinen *et al.*, 2015). Chez la poule Fayoumi, une mutation du gène *SV2A* (*Synaptic Vesicle Glycoprotein 2A)* induit le syndrome d'épilepsie réflexe photosensible sous la forme de convulsions déclenchées en réponse à une forte stimulation visuelle/lumineuse. La mutation d'un site accepteur d'épissage du gène *SV2A* ségrège dans la lignée Fepi, modèle animal pour l'étude de ce type d'épilepsie chez l'homme.

En revanche, chez les animaux de plus gros gabarit, comme les bovins et chevaux, très peu d'études épidémiologiques ont été réalisées à ce jour. Dans la majorité des cas cliniques, les convulsions ont une origine acquise environnementale, traumatique ou infectieuse déterminée (D'Angelo *et al.*, 2015, Lacombe *et al.*, 2012) et l'étiologie génétique, plus rare, est souvent peu ou pas envisagée. Néanmoins, trois syndromes épileptiques avec une forte prédisposition génétique ont été précédemment décrits : le syndrome de convulsions et ataxie familiale en race Aberdeen Angus et l'épilepsie juvénile familiale en race Brune, tous deux avec une transmission dominante supposée (Aketson *et al.*, 1944), ainsi que l'épilepsie idiopathique généralisée en race Hereford à transmission récessive (OMIA 000344-9913 ; http://www.omia.org). Seul ce dernier syndrome a fait l'objet d'une caractérisation génétique fine qui a abouti à l'identification d'une mutation récessive causale et la mise en place d'un test génétique pour cette anomalie. Cependant ni les travaux de cartographie du locus, ni le gène et la mutation en cause n'ont été publiés à ce jour.

Récemment, un nouveau syndrome de type épileptique a émergé en race Parthenaise. Il se caractérise par des crises de convulsions apparaissant chez les veaux entre un à deux mois après la naissance. La Parthenaise est une race bovine allaitante locale à fort développement musculaire. Les jeunes animaux en croissance sont considérés comme susceptibles aux carences nutritionnelles pouvant engendrer un comportement épileptique. De ce fait, les cas d'épilepsie jusqu'alors sporadiques dans cette race, n'ont pas été rapportés en tant qu'anomalie à l'Observatoire National des Anomalies Bovines (ONAB). Cependant, la persistance des crises dans le temps chez les animaux atteints, l'absence de pathologies ou de facteurs explicatifs pour les veaux ayant bénéficié d'un suivi vétérinaire et une soudaine augmentation de la fréquence des veaux épileptiques dans les naissances, ont peu à peu poussé à envisager une cause génétique. Depuis 2013, un programme d'épidémio-surveillance de l'épilepsie en race Parthenaise a été entrepris au sein de l'ONAB en partenariat avec l'Organisme de Sélection (OS) de la race afin de recenser les animaux atteints, caractériser le phénotype et confirmer l'étiologie génétique supposée pour cette nouvelle anomalie.

B. Matériel et Méthodes

1. Animaux étudiés

Depuis 2013, 58 veaux de race Parthenaise présentant des symptômes et crises de convulsions associés à un syndrome épileptique ont été répertoriés à l'ONAB, en particulier suite au signalement d'une vétérinaire, puis grâce à la participation forte de l'Organisme de Sélection (OS) de la race Parthenaise. Ces animaux proviennent de trente-huit élevages différents répartis sur huit départements français couvrant le berceau de la race et ses principales zones d'extension (79, 85, 49, 61, 71, 17, 44, 53). Ils sont issus d'accouplements entre cinquante-huit mères et trente-et-un pères différents. Deux animaux seulement ont des parents inconnus des systèmes d'information. A notre connaissance, aucun des parents n'a lui-même exprimé les symptômes de l'anomalie observés chez leurs descendants.

Des prises de sang ou des biopsies de cartilage d'oreille ont été réalisées sur les veaux épileptiques ainsi que sur quarante-quatre de leurs parents présents en ferme (38 mères et 6 pères) par un technicien de l'OS Parthenaise ou par un vétérinaire praticien lors de visites dans les élevages. Pour chacun des 102 échantillons biologiques, l'ADN génomique a été extrait à partir du sang à la plateforme d'extraction du CRB-GADIE de l'INRA (Centre de Ressources Biologiques pour la Génomique des Animaux Domestiques et d'Intérêt Economique) ou de la biopsie par un protocole d'extraction standard phénol-chloroforme.

L'ensemble des animaux atteints déclarés à l'ONAB ont été pris en compte dans l'analyse du pedigree relatif à l'anomalie. La description clinique de l'anomalie a été précisée par une enquête téléphonique rétrospective effectuée sur la moitié des élevages ayant signalé des animaux atteints. Les animaux recensés de 2013 à 2015 (39 atteints et 29 parents) ont été génotypés avec la puce Illumina Bovine SNP50 à des fins de cartographie génétique. Deux animaux atteints ont fait l'objet d'un séquençage complet de leur génome. Les cas recensés en 2016, au nombre de 17, ont été génotypés par PCR et séquençage Sanger afin de vérifier leur génotype homozygote pour la mutation candidate.

Par ailleurs, 554 animaux (126 femelles et 428 mâles) génotypés avec les puces Illumina Bovine SNP50 ou Illumina BovineHD (777K) dans le cadre d'un autre projet de recherche mené dans l'équipe (projet GEMBAL), ont été utilisés comme population de contrôle.

2. Analyse des pedigrees

L'analyse des pedigrees, le calcul des contributions génétiques et la recherche d'un ou plusieurs ancêtre(s) commun(s) aux individus atteints ont été effectués à l'aide du logiciel PEDIG (Boichard 2002).

3. Cartographie de l'anomalie

Après application de contrôles qualité standards (vérification de la compatibilité des génotypes des produits avec ceux de leurs parents, *call rate* par animal > 0,95, respect de l'équilibre de Hardy-Weinberg), les génotypes manquants ont été imputés et les phases ont été reconstituées pour l'ensemble des individus (atteints, parents et contrôles) avec le logiciel FIMPUTE (Sargolzaei, Chesnais, and Schenkel 2014). Les cas d'épilepsie étant rapportés chez des animaux mâles et femelles en proportions égales, les marqueurs des chromosomes X et Y n'ont pas été conservés pour analyse. Au final 43 801 marqueurs répartis sur les 29 autosomes ont été utilisés.

La localisation de l'anomalie a été réalisée par cartographie par homozygotie, qui consiste à rechercher les régions homozygotes identiques entre les individus atteints (N=36) par comparaison avec une population contrôle d'individus sains (N=580, parents et individus parthenais génotypés). Le logiciel utilisé HOMAP (développé en interne dans l'équipe) est un dérivé du logiciel ASSHOM et s'appuie sur la statistique de test décrite par Charlier *et al.*, (2008).

4. Séquençage, alignement des génomes, détection et annotation des variants

Les génomes de deux animaux épileptiques ont été séquencés à la plateforme génomique Get-PlaGe de Toulouse (http://genomique.genotoul.fr/). Ils ont été choisis parmi les animaux confirmés homozygotes pour l'haplotype associé à l'anomalie de façon à minimiser la taille de la région commune. Pour chaque animal, des librairies paired-end de fragments d'ADN de 300 pb ont été générées, puis poolées et séquencées sur une même piste de la plateforme de séquençage HiSeq 3000 d'Illumina. Les lectures de 150 pb en moyenne ont ensuite été alignées sur le génome de référence bovin UMD3.1 à l'aide du logiciel BWA (Li and Durbin 2009). La couverture finale moyenne obtenue pour chaque génome est de 11.62 et 13.9 x, après retrait des dupliquas de PCR.

Les SNP (Single nucleotide polymorphism) et petites insertions et délétions (InDels) ont été détectés avec l'outil mpileup de SAMtools (Li *et al.*, 2009) puis annotés avec le logiciel Variant Effect Predictor d'Ensembl (McLaren *et al.*, 2016). Dans l'intervalle d'homozygotie, les variations structurales (grandes insertions, délétions, inversions ou duplications) ont été recherchées avec le logiciel PINDEL (Ye *et al.*, 2009), puis visuellement par fenêtres de 10 kb avec le logiciel de visualisation graphique de données de séquences *Integrative Genomics Viewer* (IGV) (Robinson *et al.*, 2011).

5. Filtres appliqués aux variants et identification de la mutation causale

Afin d'isoler les variants candidats potentiels, nous avons filtré les données de séquence pour retenir ceux (i) situés dans l'intervalle d'homozygotie identifié (chr24:41289711-42467023), (ii) présents à l'état homozygote chez les deux individus atteints séquencés et (iii) et absents des données de séquence de 1323 individus contrôles. Ces derniers sont issus du rassemblement des données WGS du consortium

1000 génomes bovins (N=1147 ; Daetwyler *et al.*, 2014) et des données WGS disponibles au sein de l'équipe (N=272 dont 96 individus en commun avec le projet précédent ; Boussaha *et al.*, submitted). Aucun de ces animaux n'étant de race Parthenaise, nous les avons considérés sains et non-porteurs de cette anomalie. A l'issue de cette étape, les variations restantes ont été évaluées en fonction de leur annotation fonctionnelle. Une priorité a été donnée aux variants prédits avec un effet délétère potentiel sur une protéine : gain ou perte d'un codon stop, décalage du cadre de lecture, modification des sites d'épissage et les substitutions d'acide aminé prédites délétères par SIFT (Kumar *et al.*, 2009).

6. Conservation entre espèces et prédiction de l'effet de la mutation

L'étude de la conservation du site d'épissage affecté a été réalisée à partir de l'alignement des séquences nucléotidiques des gènes orthologues au gène *MTCL1* bovin de 39 espèces de vertébrés disponible dans la base de données Ensembl (www.ensembl.org).

Une prédiction *in silico* de l'effet de la mutation au niveau du site accepteur d'épissage a été réalisée à l'aide du logiciel CRYP-SKIP accessible en ligne (serveur en ligne : http://cryp-skip.img.cas.cz ; Divina *et al.*, 2009). A partir de la séquence d'ADN génomique, le logiciel estime les probabilités de deux mécanismes les plus fréquents liés à une modification d'un site accepteur d'épissage : la suppression totale de l'exon lors de l'épissage ou une activation d'un site accepteur cryptique en aval du site réel. La séquence d'ADN génomique bovine comprenant une partie de l'intron 7 et l'exon 8 du gène *MTCL1* a été extraite à partir l'UMD3.1 depuis la base de données Ensembl (www.ensembl.org). Dans un deuxième temps, des prédictions *in silico* de la protéine produite en fonction des résultats de CRYP-SKIP a été réalisée à partir des séquences cDNA des 14 exons de *MTCL1* extraites depuis Ensembl (www.ensembl.org) et traduites en séquences protéiques avec le logiciel en ligne *Sequence Manipulation Suite* (http://www.bioinformatics.org/sms2). Enfin, les séquences résultantes ont été alignées entre elles à l'aide de du logiciel *Clustal Omega* (http://www.ebi.ac.uk/Tools/msa/clustalo ; (Sievers *et al.*, 2011).

7. Confirmation de la mutation candidate dans MTCL1 par PCR et séquençage Sanger

Pour confirmation, deux trios père-mère-veau porteurs de l'haplotype sur marqueurs de la puce SNP50K, les trois animaux discordants identifiés lors de la cartographie et les 17 nouveaux cas épileptiques collectés fin 2015 - début 2016 ont été génotypés pour le polymorphisme candidat chr24 g.41661691G>A par PCR et séquencage Sanger. Le couple d'amorces, MTCL1_F:TGCTTCAAGATAGCCATGACC et MTCL1_R: CGTGCCCTTACTATGTCCTCA, a été défini sur l'assemblage du génome bovin UMD3.1 à l'aide du logiciel Primer3 (Rozen et Skaletsky, 2000). L'amplification par PCR a été réalisée sur un thermocycler Eppendorf Mastercycler pro avec la PCR-polymerase Go-Taq Flexi de Promega, en respectant les instructions d'utilisation du fournisseur.

Les produits de PCR ont été purifiés puis séquencés selon la méthode de séquençage Sanger conventionnelle par Eurofins MWG (Allemagne). Enfin, les séquences obtenues ont été analysées à l'aide du logiciel novoSNP pour la détection des variants (Weckx *et al.*, 2005).

C. Résultats et Discussion

1. *Emergence d'un syndrome d'épilepsie généralisée récessive en race bovine Parthenaise*

Avec la participation active des techniciens de l'organisme de sélection de la race Parthenaise, 58 animaux (28 mâles et 30 femelles) supposés épileptiques et nés entre 2011 et 2016 (Figure 3) ont été rapportés à l'ONAB depuis 2013. Le terme « épilepsie » a été associé dans un premier temps par éleveurs et techniciens pour décrire ces veaux manifestant des crises convulsives sans facteur environnemental évident et similaires aux descriptions de crises épileptiques chez l'homme.

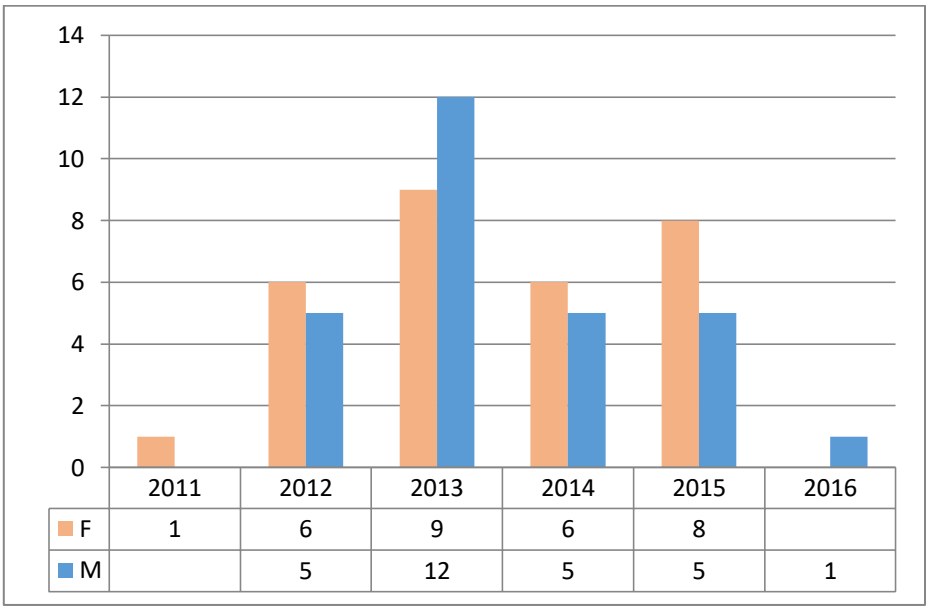

Figure 3 : Veaux épileptiques déclarés à l'ONAB sur la période 2013-2016, répartis en fonction de leur année de naissance

Par le biais d'une enquête dans les élevages ayant recensé au moins un animal atteint, le phénotype clinique de l'anomalie a pu être précisé (Anne Relun, communication personnelle). Les premières crises apparaissent chez les jeunes veaux âgés d'un à deux mois. Elles se déroulent en trois temps. Tout d'abord, l'animal s'isole de ses congénères, ses membres se raidissent, puis il tombe sur le flanc. Il entre alors dans une phase convulsive à terre avec une conservation partielle de la conscience (existence d'une faible réaction aux stimuli externes). L'animal, en décubitus latéral, présente des tremblements et

contractions incontrôlées des muscles avec des membres raides, un cou tendu vers l'arrière, et des yeux révulsés (Figure 4A). En fin de crise, il conserve des membres raides avec au relevé une démarche chancelante, désorientée et ataxique (Figure 4B). Ces derniers symptômes se dissipent en 5 à 10 min après le relevé pour revenir à un comportement normal sans séquelles.

Figure 4 : Veaux parthenais épileptiques observés en cours (A) et en fin de crise (B).
(source: Anne Relun ; OS Parthenaise)

Selon les observations des éleveurs, les durées et fréquences des épisodes sont assez variables : trois quarts d'heure en moyenne (de 15 min à plus d'une heure), une à deux fois par semaine (entre une crise par jour et une crise tous les deux mois). En dehors des crises, les animaux ne présentent pas d'autres symptômes et se développent normalement. Les crises persistent à l'âge adulte, sans dégradation ni amélioration de l'état des animaux. Cette anomalie n'est donc pas directement létale et les quelques cas de mortalité des animaux atteints liés aux crises sont d'ordre accidentels (pendaison et étouffement par retournement).

D'autre part, nous n'avons pas mis en évidence d'anomalies de développement des organes en particulier au niveau du cerveau, de la moelle épinière ou du système cardio-vasculaire chez une génisse épileptique abattue à l'âge de trois ans. Les analyses histo-pathologiques des tissus cérébraux, nerveux et musculaires de cet animal n'ont pas révélé de lésions structurelles ou dégénératives (Anne Relun, communication personnelle).

Selon la définition donnée par la ligue internationale contre l'épilepsie et différents rapports épidémiologiques vétérinaires, les symptômes cliniques observés chez ces animaux sont compatibles avec une épilepsie. Le déroulement d'une crise, en trois temps avec une phase convulsive accompagnée d'une perte de conscience est caractéristique des crises épileptiques généralisées. L'atteinte bilatérale traduit en général une extension de la décharge électrique anormale à l'origine de la crise aux deux hémisphères cérébraux (Berendt *et al.*, 2015). Un électroencéphalogramme sera cependant nécessaire

pour confirmer les troubles de l'activité électrique neuronale et l'origine des crises de convulsions chez les animaux atteints. Nous prévoyons également de suivre l'abattage d'autres animaux atteints pour confirmer les observations réalisées sur ce premier animal.

Les animaux épileptiques de cette étude sont en proportions mâles/femelles équivalentes et issus d'accouplement entre 58 mères et 31 taureaux connus. Les pères ont entre 1 et 3 descendants atteints, sauf pour un taureau qui possède 17 veaux atteints connus de l'ONAB. Plusieurs ancêtres communs ont été identifiés dans le pedigree en particulier le taureau ECUSSON (1989) avec une contribution génétique de 9,3% parmi les animaux atteints contre 4% de contribution moyenne à l'ensemble de la population Parthenaise et son grand-père maternel MAGICIEN (1976), qui est un ancêtre commun à l'ensemble des cas. Sur les voies maternelle et paternelle, c'est la mère de MAGICIEN qui relie l'ensemble des veaux atteints entre eux.

La multiplicité des élevages touchés, l'absence de dégradation de l'état général de l'animal, l'absence de symptômes chez les parents et la présence d'ancêtres communs dans les pedigrees sont autant d'éléments en faveur d'une anomalie génétique à déterminisme récessif. Etonnamment, on retrouve une très grande similitude entre les symptômes décrits chez la Parthenaise et le syndrome d'épilepsie idiopathique qui ségrége en race Hereford (Tableau 2).

1. *La cartographie par homozygotie localise l'anomalie dans un intervalle de 1,2 Mb sur le chromosome 24*

Les études cliniques et généalogiques suggérant un déterminisme autosomal récessif, nous avons appliqué une approche de cartographie par homozygotie. L'analyse a été réalisée en utilisant les marqueurs de la puce Illumina Bovine SNP50 sur les 29 autosomes selon le modèle décrit par Charlier *et al.,* (2008). Nous avons comparé les génotypes de 36 individus atteints à un groupe contrôle de 580 individus sains, comprenant les parents typés dans le cadre de ce projet et du reste des individus de la population Parthenaise génotypés à ce jour.

Cette analyse a mis en évidence une association significative avec plusieurs dizaines de SNP localisés entre 40 et 45 mégabases sur le chromosome 24. En étudiant les génotypes à SNP des animaux atteints et des animaux contrôles, nous avons identifié un unique haplotype de 22 SNP consécutifs (ARS-BFGL-NGS-70555, Chr24 :41289711 ; au marqueur ARS-BFGL-NGS-79945, Chr24 :42467023) homozygote et commun à 33 des 36 individus atteints de l'analyse. Ce résultat confirme le déterminisme récessif simple supposé et réduit la localisation de la mutation causale à un segment de 1,2 Mb sur le chromosome 24 (chr24:41220225-42506011). En dehors des individus atteints, l'haplotype n'a jamais été retrouvé à l'état homozygote dans la population de contrôle. Sa fréquence parmi les individus sains (N=580) est estimée autour de 6% (parents des animaux atteints compris).

Tableau 2: Description clinique des symptômes et comparaison avec d'autres syndromes épileptiques chez le bovin

	Epilepsie en race Parthenaise	**Epilepsie idiopathique généralisée Hereford (www.omia.org)**	**Brune (Aketson *et al.*, 1944)**
Age à la première crise	Entre 1 et 2 mois	Variable : de la naissance à plusieurs mois	6 mois
Crises de convulsion	Généralisées Raideur musculaire Décubitus latéral Spasmes/convulsions de type tonic-clonique Perte de conscience partielle	Généralisées Décubitus latéral Raideur des membres Tremblements	Tête basse Phénomène de « mâchage » Ecume au niveau de la gueule Perte de conscience
Durée crise (totale)	15 min à 1h30	Quelques minutes à > 1 h	NA
Fréquence	~2 par semaines [1 fois par jour - 1 fois tous les 2 mois]	Variable	Variable
Facteurs déclenchants	Stress : stimulation, mouvement de troupeau	Stress environnementaux : changement température extérieure, activité physique	Excitation
Comportement entre les crises	Normal	Normal	Normal
Croissance	Normale	Normale	NA
Dégradation dans le temps	Non	Non précisé	Perte de tonus dans les membres postérieurs
Régression dans le temps	Non	Non précisé	Diminution de fréquence, disparition entre 1 à 2 ans
Mortalité	Non (sauf accidents)	Non (sauf accidents)	Non
Déterminisme	Autosomal récessif	Autosomal récessif Gènes et mutation causale identifiés, non publiés	Autosomal dominant

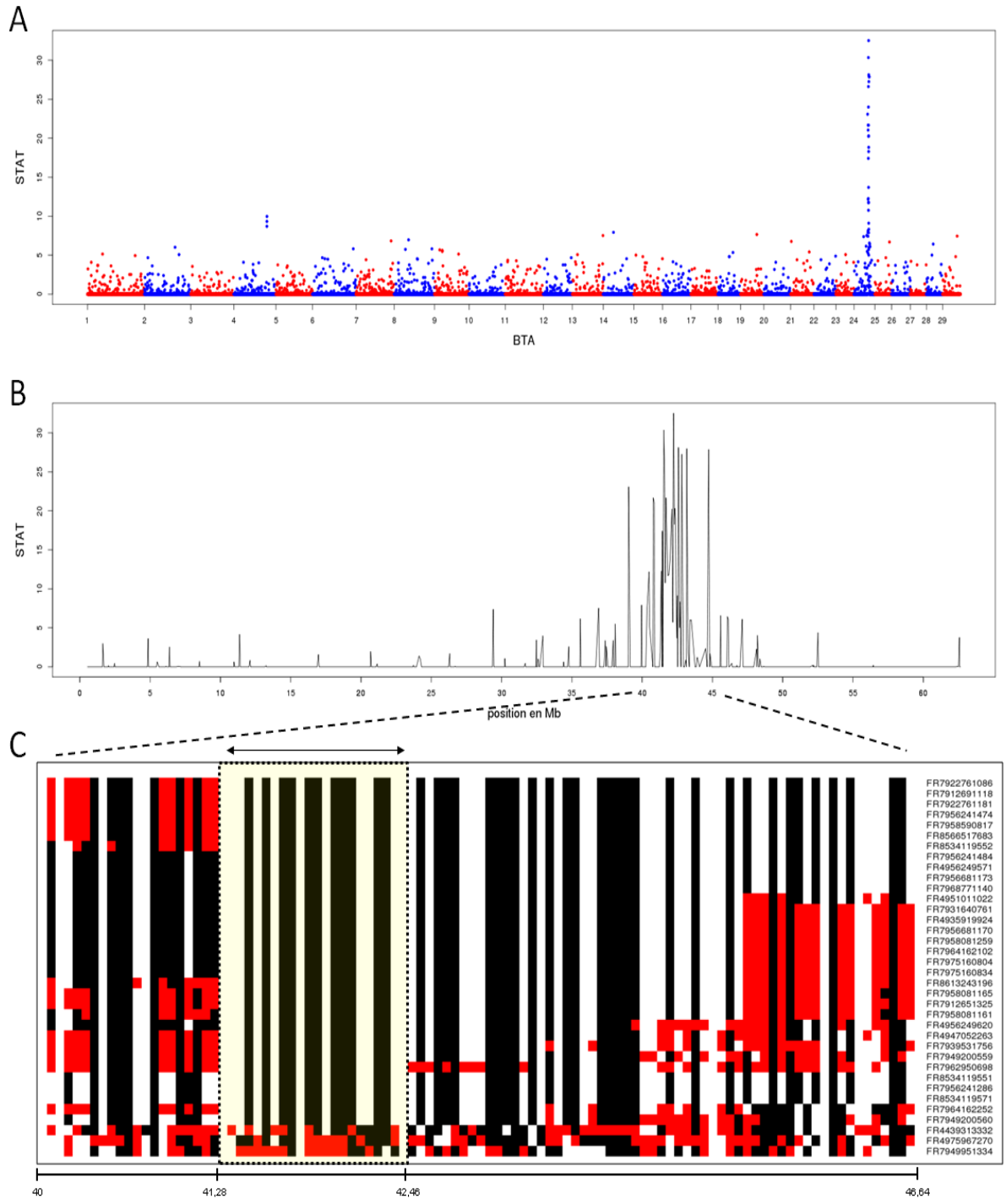

Figure 5 : Cartographie par homozygotie de la mutation responsable du syndrome d'épilepsie en race Parthenaise

A. Manhattan plot des résultats de cartographie par homozygotie opposant 36 animaux supposés atteints aux 580 individus Parthenais contrôles ; B. Zoom sur le chromosome 24 ; C. Comparaison des génotypes des 36 animaux supposés atteints pour les 100 marqueurs du BTA 24 (100 marqueurs de 40Mb à 46,6 Mb) au niveau de l'intervalle d'homozygotie détecté. Pour chaque SNP, les génotypes homozygotes sont représentés en noir ou blanc et les génotypes hétérozygotes en rouge. L'intervalle d'homozygotie est indiqué par les pointillés et la double flèche.

2. *Un polymorphisme touchant un site accepteur d'épissage dans le gène MTCL1 est identifié comme mutation candidate responsable de l'épilepsie*

L'intervalle de 1,2 Mb identifié sur le chromosome 24 contient 11 gènes différents (Tableau 3). Parmi eux, seul le gène *NDUFV2* (*nadh-ubiquinone oxido-reductase flavoprotein 2)* qui code pour une sous-unité d'un complexe protéique de la chaine respiratoire mitochondriale, est associé chez l'homme au syndrome de déficit en complexe mitochondrial 1 (OMIM 252010) pour lequel les crises épileptiques sont un symptôme. Cependant, le phénotype général de ce syndrome, qui inclut une hypertrophie cardiaque et une encéphalopathie congénitales, n'est pas compatible avec celui de animaux de cette étude (Benit *et al.*, 2003). De plus, aucun de ces 11 gènes n'a été associé à un syndrome épileptique chez d'autres espèces animales modèles ou domestiques.

Tableau 3 : Gènes localisés dans l'intervalle d'homozygotie de 1,2 Mb sur le chromosome 24

Ensembl Gene ID	Gène	Début (pb)	Fin (pb)	Syndrome Humain (OMIM)
ENSBTAG00000024015	PTPRM	40768713	41437121	
ENSBTAG00000010356	RAB12	41532521	41541737	
ENSBTAG00000009459	MTCL1	41576970	41689810	
ENSBTAG00000004871	NDUFV2	41855938	41874838	*MITOCHONDRIAL COMPLEX 1 DEFICIENCY* [OMIM:252010]
ENSBTAG00000002755	ANKRD12	41913310	41975447	
ENSBTAG00000001805	TWSG1	41986506	42008147	
ENSBTAG00000021250	RALBP1	42061941	42079438	
ENSBTAG00000009822	PPP4R1	42093001	42121847	
ENSBTAG00000046533	RAB31	42256081	42305185	
ENSBTAG00000017276	TXNDC2	42311245	42313446	
ENSBTAG00000017279	VAPA	42332694	42368659	

Devant l'absence de gène candidat évident, nous avons adopté une approche d'analyse des séquences de génomes complets de deux individus atteints, homozygotes pour l'haplotype associé. Dans l'intervalle d'homozygotie, nous avons identifié 1269 SNP et petits InDel homozygotes pour l'allèle alternatif chez les deux parthenais épileptiques. L'analyse de la région avec PinDel puis par visualisation avec IGV n'a mis en évidence aucune variation structurale additionnelle. Après filtre contre les données de séquence de 1323 individus sains, nous avons conservé seulement neuf variants (huit SNPs et un InDel) trouvés uniquement chez les deux individus atteints (Tableau 4).

Seule la substitution g.41661691G>A est supposée avoir un effet délétère fort sur le gène dans lequel elle est située. En effet, elle modifie le motif -AG- du site accepteur d'épissage entre l'intron 7 et l'exon 8 du gène *Microtubule Crosslinking Factor 1* (*MTCL1*). En complément, nous avons étudié la conservation de l'ensemble de ces nucléotides chez les vertébrés et seul le variant g.41661691G>A est parfaitement conservé (Figure 6).

Cette observation suggère une contrainte évolutive forte au niveau de ce site d'épissage. Le polymorphisme g.41661691G>A constitue donc la seule mutation candidate délétère dans l'intervalle de localisation de l'anomalie.

Tableau 4 : variants candidats après analyse des données de séquence

Chr	Position (pb)	Ref	Alt	Conséquence	Gène	Numéro d'accession Ensembl
24	41661691	G	A	Site accepteur d'épissage	MTCL1	ENSBTAG00000009459
24	41730176	G	A	Intergénique		
24	41969018	T	C	Intronique	ANKRD12	ENSBTAG00000002755
24	42106354	T	C	Intronique	PPP4R1	ENSBTAG00000009822
24	42112442	CAGAGGGAAGA	CAGA	Intronique	PPP4R1	ENSBTAG00000009822
24	42376980	A	G	Intergénique		
24	42380420	G	A	Intergénique		
24	42466115	C	T	Intergénique		
24	42466136	C	A	Intergénique		

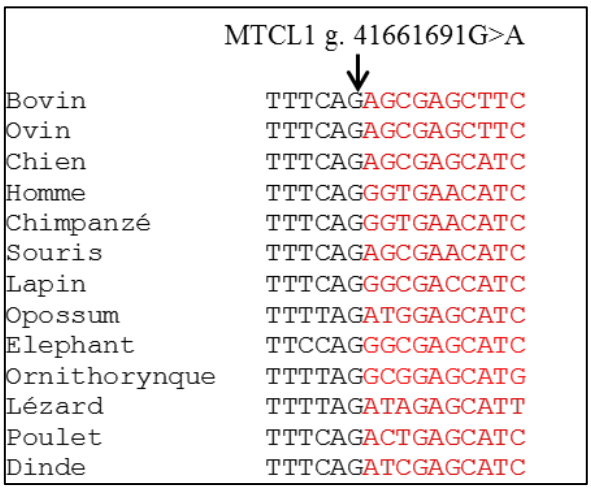

Figure 6 : Conservation de la séquence nucléotidique du gène MTCL1 entre vertébrés au niveau de la mutation g.41661691G>A.

Le gène *MTCL1* code pour une nouvelle protéine en interaction avec les microtubules, dont la fonction précise reste peu décrite à ce jour. Les microtubules sont des éléments dynamiques du cytosquelette indispensables dans de nombreux processus et fonctions cellulaires. Dans les neurones, ils interviennent

dans la migration des cellules nerveuses, la formation et la plasticité des axones, la mise en place de leur polarité et le maintien de leur morphologie spécifique. Ils contribuent également à l'ensemble des transports synaptiques nécessaires à l'activité bioélectrique inhérente au fonctionnent neuronal (Sakakibara *et al.,* 2013). En conséquence, les microtubules sont des cibles d'étude privilégiées dans plusieurs maladies neuro-dégénératives (Dubey *et al.,* 2015) et leur rôle dans les phénomènes électriques à l'origine des crises épileptiques a récemment été démontré. En effet, chez le rat, un lien a été établi entre une altération de l'organisation et de la dynamique des microtubules et une hyperexcitabilité électrique des neurones (Carletti *et al.,* 2016 ; Xu *et al.,* 2016).

Pour en revenir à MTCL1, des études récentes ont démontré son rôle dans les processus d'organisation et de régulation de la dynamique des microtubules au sein des cellules différenciées et polarisées, comme les cellules épithéliales ou les neurones. En particulier, l'inactivation de ce gène empêche le développement de faisceaux de microtubules normaux et fonctionnels dans des cellules épithéliales en culture (Sato *et al.,* 2013 ; Sato *et al.,* 2014).

Nous pouvons envisager différentes hypothèses d'impact de la mutation g.41661691G>A sur la protéine MTCL1: l'activation d'un autre site accepteur, la supression de l'exon 8 lors de l'épissage, ou une dégradation de l'ARN muté par les processus cellulaires (*RNA decay*), ce qui conduirait dans ce cas à l'absence de protéine chez les animaux homozygotes. Dans les deux premières situations, nous avons tenté de prédire l'effet de la subtitution g. 41661691G>A sur l'épissage du gène à l'aide du logiciel CRYP-SKIP (Divina *et al.,* 2009), un logiciel estimant les positions de sites cachés et la probabilité de leur activation par rapport à celle de la supression de l'exon. La probabilité de 0,39 obtenue est légèrement en faveur d'une délétion de l'exon 8 lors de l'épissage (Figure 7). Deux sites accepteurs probables sont proposés dans l'exon 8. Le premier site (CRYP1) est celui qui affecterait le plus fortement la protéine en générant un changement de cadre de lecture à partir de l'acide aminé 1020 et la terminaison prématurée au deux-tiers de la protéine. Ceci entrainerait la suppression du domaine de liaison avec la protéine MARK2 (*Microtubule Affinity Regulating Kinase 2*) et du domaine KR-rich importants pour les interactions avec les microtubules (Sato *et al.,* 2013 ; Sato *et al.,* 2014). Le second site testé aurait un effet moindre sur la protéine avec simplement une délétion des 43 premiers acides aminés de l'exon 8.

Nous n'avons pas eu le temps à ce jour de vérifier expérimentalement l'effet de la mutation sur l'expression de la protéine MTCL1. Il faudrait extraire l'ARN de prélèvements d'encéphale d'animaux homozygotes atteints et de contrôles, générer les ADNc et construire un test par PCR permettant de vérifier la présence ou l'absence de l'exon 8. Celui–ci étant relativement court, nous pouvons envisager par exemple d'utiliser trois amorces (la première dans l'exon 7 puis les deux amorces antisens complémentaires dans l'exon 8 et dans l'exon 9) et de faire migrer les amplicons par électrophorèse sur gel d'agarose 3% afin de révéler des différences de tailles éventuelles.

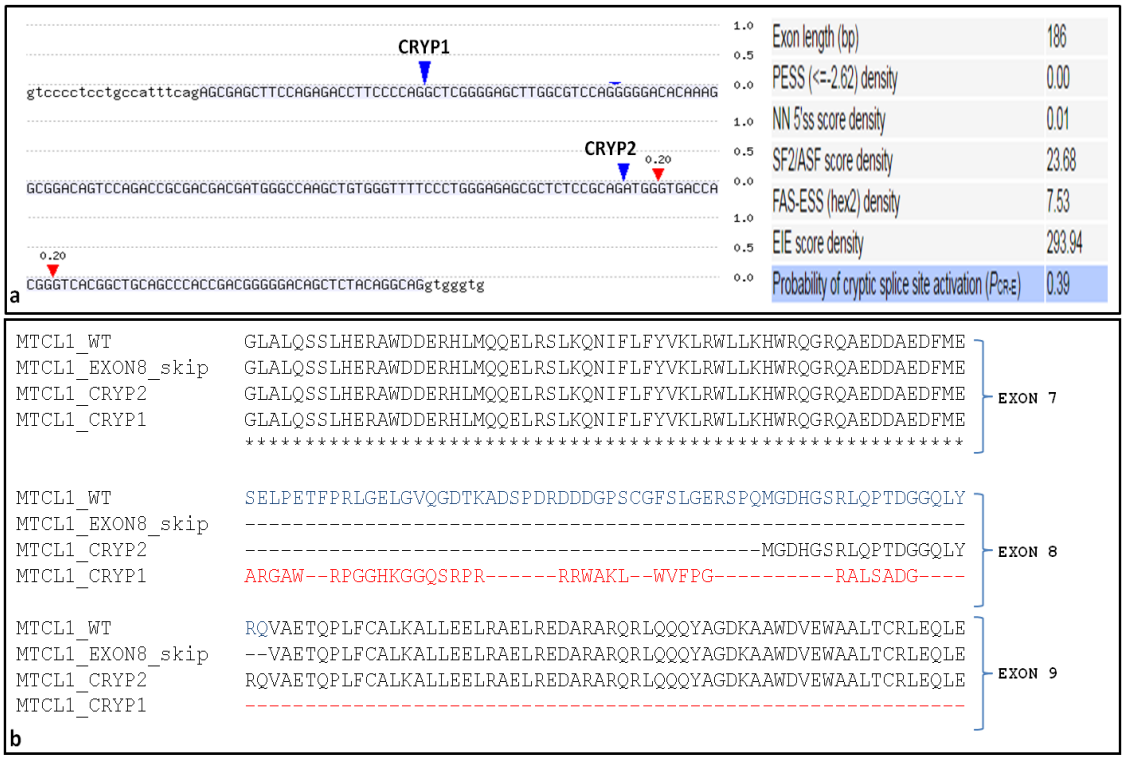

Figure 7: Prédiction in silico de l'effet de la mutation g.41661691G>A sur la protéine MTCL1
a. Copie d'écran des résultats du logiciel CRYP-SKYP indiquant les sites accepteurs (flèches bleues) et donneurs (flèches rouges) d'épissage cachés au niveau de l'intron 7 et l'exon 8 de la séquence du gène MTCL1. **b**. Comparaison des différentes séquences protéiques de MTCL1 obtenues en fonction des trois hypothèses d'effet de la mutation sur le site accepteur : la suppression de l'exon 8 (MTCL1_EXON8_skip) ou l'activation d'un autre site accepteur d'épissage (MTCL1_CRYP1 et MTCL1_CRYP2). Dans le cas de CRYP2, la protéine serait réduite d'une partie de l'exon 8 (43 acides aminés), tandis que le site CRYP1 entrainerait un décalage dans le cadre de lecture de l'exon 8 et la terminaison de la protéine.

3. Confirmation de l'existence de la mutation chez les animaux épileptiques

Nous avons génotypé par PCR et séquençage Sanger les deux individus atteints dont nous disposions de la séquence génomique, deux autres individus atteints et chacun de leurs parents, eux-mêmes porteurs de l'haplotype 50K associé à l'anomalie. En complément, nous avons également génotypé les trois individus déclarés atteints mais non porteurs de l'haplotype 50K ainsi que les nouveaux cas de veaux épileptiques déclarés à l'ONAB au cours de l'hiver 2015-2016 (N=17).

Les individus atteints homozygotes de l'haplotype sont bien homozygotes pour l'allèle muté. Leurs parents sont porteurs hétérozygotes. Les trois animaux identifiés non porteurs de l'haplotype lors de la cartographie sont non porteurs de la mutation (ie. homozygotes pour l'allèle de référence). Parmi les 17 nouveaux cas déclarés, 14 sont homozygotes pour la mutation et trois sont homozygotes pour l'allèle de référence, ce qui porte le nombre de discordants à 6 individus.

Jusqu'à aujourd'hui, le recrutement des animaux épileptiques en race Parthenaise a été majoritairement réalisé par déclaration spontanée des éleveurs et techniciens de l'OS à l'ONAB. Nous avons établi une

première description clinique sur un tableau rétrospectif auprès de ces éleveurs mais aucun des animaux déclarés épileptiques n'a fait l'objet d'examens cliniques complémentaires (biochimie sanguine, éléments infectieux …). De plus, les symptômes évoqués peuvent conserver une part d'incertitude due au temps écoulé entre les observations et leur déclaration, ce qui rend difficile l'exclusion des cas hétérogènes dont la cause serait non génétique. En effet, chez le bovin, des symptômes convulsifs sont souvent provoqués par des lésions du système nerveux après un accident, en réaction à un agent toxique ou infectieux, ou bien d'origine métabolique avec par exemple un déséquilibre de la balance en ions minéraux (ie. calcium et magnésium), nécessaires au fonctionnement du système neuromusculaire (D'Angelo *et al.*, 2015). L'existence de phénocopies dans les cas recensés n'est donc pas impossible. Cette hypothèse expliquerait les six animaux supposés épileptiques sur la base d'une observation phénotypique similaire mais qui sont non-porteurs de l'haplotype et de la mutation candidate.

Au travers de l'enquête téléphonique, nous avons tenté de re-confirmer ou infirmer auprès des éleveurs le phénotype de ces animaux aux génotypes discordants pour la mutation candidate. Les premiers retours indiqueraient quelques différences minimes et difficilement détectables au niveau de la durée des crises, l'âge d'apparition et la fréquence dans le temps. L'un d'entre eux aurait peut-être pu souffrir d'anoxie à la naissance (vêlage très difficile sans césarienne). Un autre aurait pu chuter au cours d'une manipulation de pesée vers 3 mois et n'avoir exprimé des crises qu'après cet évènement. Cependant, ce ne sont que des hypothèses et d'autres analyses cliniques (suivi dans le temps, récurrence des crises…) seront nécessaires pour conclure sur la situation de ces individus discordants.

D. Conclusion

Avec cette étude nous rapportons un nouveau syndrome récessif d'épilepsie idiopathique généralisée qui ségrége en race Parthenaise, constituant la seconde épilepsie récessive avec celle de la race Hereford caractérisée chez le bovin. Par la cartographie par homozygotie sur une trentaine de cas et le séquençage du génome complet de deux atteints, nous avons localisé l'anomalie dans un intervalle de 1,2 Mb sur le chromosome 24 et proposons un polymorphisme modifiant un site accepteur d'épissage du gène *MTCL1* comme mutation candidate. Avec la grande similitude observée entre les phénotypes, il serait intéressant de pouvoir comparer le gène mis en cause en race Hereford avec notre gène candidat, dans l'éventualité où ils seraient identiques ou appartenant aux mêmes voies ou familles de gènes.

Bien que l'épilepsie du veau Parthenais n'ait été déclarée à l'ONAB qu'à partir de 2013, l'apparition de cas sporadiques dans les élevages aurait sans doute commencé à partir des années 2000. Les symptômes évoquent au premier abord d'autres étiologies et la cause génétique n'a été envisagée que récemment avec l'augmentation de la fréquence de l'émergence d'animaux épileptiques.

L'utilisation massive d'un taureau porteur de l'anomalie ayant par ailleurs un très bon index, en particulier en facilités de vêlage, a sans doute induit une dissémination forte de l'anomalie ces dernières années et est probablement à l'origine de l'augmentation des cas dans les naissances.

L'épilepsie constitue la première anomalie génétique connue de la race Parthenaise. Elle n'est pas létale et impacte peu la production. Cependant, ce syndrome induit des animaux plus « fragiles » et qu'il faut surveiller, ce qui augmente la charge de travail de l'éleveur. Elle augmente aussi les risques de perte de l'animal au cours du transport et à l'abattage (risques de blessures ou encore saisie des carcasses « noires » des animaux qui ont fait une crise peu de temps avant l'abattage). Les crises épileptiques donnent également une mauvaise image concernant la santé et le bien-être de ces veaux. Les symptômes nerveux engendrent également une peur par rapport aux crises sanitaires connues du grand public (ex. ESB). Les tests sur haplotype, puis sur mutation causale (après validation fonctionnelle de celle-ci) permettront de mettre en place rapidement un contrôle de cette anomalie dans la race en limitant les taureaux reproducteurs porteurs ainsi que les croisements donnant une naissance à risque.

A notre connaissance, notre étude est la première qui associe directement une mutation du gène *MTCL1* à un syndrome épileptique. Si nous confirmons l'effet, notre étude pourrait apporter des pistes supplémentaires pour la compréhension des syndromes épileptiques chez l'homme.

Les travaux réalisés sur cette anomalie sont encore en cours. Plusieurs points restent à préciser dans la description fine du phénotype de l'anomalie et il nous reste à valider l'effet de la mutation candidate avec des analyses d'expression du gène. Il est envisagé de faire hospitaliser quelques animaux de façon à pouvoir réaliser des examens cliniques complémentaires recommandés dans les diagnostiques d'épilepsie ; en particulier un électroencéphalogramme qui est difficilement réalisable en ferme. Nous travaillons avec l'OS parthenaise et les éleveurs pour recenser les animaux épileptiques prêts à être abattus pour commercialisation bouchère, afin de récupérer les prélèvements de tissus nécessaires. C'est une étape assez compliquée car les animaux sont valorisables par l'éleveur (les mâles autour de 15 mois et les femelles vers 36 mois). Il n'est pas envisageable d'acheter des veaux et les étudier en école vétérinaire ou station de recherche. Nous avons à ce jour deux cas et une campagne de prélèvements est prévue fin 2016/début 2017. Nous devrions être en mesure d'apporter une confirmation de l'effet de la mutation par l'analyse de l'ARN de ces animaux d'ici la fin de l'année 2017. La fonction du gène étant peu connue, nous envisageons de réaliser une étude RNA-seq pour approfondir le mécanisme moléculaire sous-jacent et comprendre en quoi le gène *MTCL1* serait lié à l'épilepsie.

II. Bilan du chapitre

Dans ce chapitre, nous avons présenté l'étude d'une anomalie génétique selon l'approche classique du phénotype au génotype en utilisant les outils technologiques les plus récents, ce qui a permis d'aboutir de manière efficace à l'identification d'une mutation candidate. Elle montre encore une fois la puissance apportée par le génotypage sur puce à SNP dans les approches de cartographie par homozygotie, mais aussi la puissance apportée avec l'utilisation des données de séquences de génome complet. D'une part, par la puissance du filtre disponible pour la détection du polymorphisme délétère en passant d'un millier de variants à un seul candidat délétère dans la région étudiée. Cette efficacité est liée au nombre de données de séquence qui nous servent de contrôle, dans notre cas plusieurs centaines d'individus de races différentes grâce surtout au consortium « *1000 bull genomes* ». Par ailleurs, comme aucun gène ne constituait un bon candidat fonctionnel, il aurait été nécessaire de séquencer les 11 gènes de l'intervalle pour identifier le polymorphisme responsable.

CHAPITRE 3 : CARTOGRAPHIE ET IDENTIFICATION DE MUTATIONS RECESSIVES A PARTIR D'UN SEUL GENOME : SYNDROME D'EPIDERMOLYSE BULLEUSE JONCTIONNELLE EN RACE CHAROLAISE.

Comme nous l'avons vu dans l'introduction et le chapitre précédent, la centralisation de la collecte des cas au sein des observatoires nationaux, la cartographie génétique avec des puces haute-densité et le séquençage complet du génome d'individus atteints sont des approches très efficaces pour identifier les mutations causales et permettre leur gestion rapide dans les populations bovines. Elles permettent notamment d'obtenir des résultats avec un nombre restreint d'animaux atteints (typiquement 5 à 10) et ce même s'ils sont très consanguins. C'est l'atout majeur de cette approche, mis en avant par l'étude de Charlier *et al.,* (2008), qui fait référence dans le monde bovin avec la cartographie génétique de cinq anomalies récessives disposant de seulement 3 à 12 individus atteints chacune.

Ces méthodes sont efficaces, mais par principe la cartographie par homozygotie nécessite la comparaison des génotypes de plusieurs individus atteints, c'est-à-dire au moins deux. Malgré les dispositifs de surveillance et la communication mise en place par les observatoires, il reste parfois difficile d'arriver à temps pour prélever le matériel biologique nécessaire aux analyses. C'est d'autant plus difficile lorsque l'anomalie entraîne une mortalité périnatale ou demande une euthanasie rapide des animaux, non viables en élevage. Les cas sont alors déclarés a posteriori sans que les prélèvements et autopsies n'aient pu être effectués. Si l'on ajoute à cela la fréquence généralement faible d'une anomalie émergente dans la population, il peut s'écouler plusieurs années avant de disposer d'un nombre suffisant de cas pour commencer une étude.

Nous avons fait face à ce type situation pour le syndrome d'épidermolyse bulleuse jonctionnelle (EBJ), ségrégant à très faible fréquence dans la race Charolaise. L'EBJ est une anomalie congénitale engendrée par un défaut des protéines au niveau de la jonction dermo-épidermique entraînant une grande fragilité mécanique de la peau. Les veaux atteints naissent avec de grandes lésions cutanées à vif et sont généralement euthanasiés rapidement après la naissance. C'est une anomalie présumée récessive en race Charolaise, connue depuis le milieu des années 1980, et pour laquelle un cas a fait l'objet d'une description clinique dans le milieu des années 2000 (Guaguere *et al.*, 2004). Pour cette anomalie, seulement quatre veaux atteints avaient été déclarés à l'ONAB, dont deux sans matériel biologique, ce qui a limité la mise en place d'une étude par cartographie classique.

En s'appuyant sur les nouvelles technologies de séquençage, nous avons développé une stratégie alternative pour identifier la mutation causale, une grande délétion dans le gène de l'intégrine beta 4. Celle stratégie s'appuie directement sur les données de séquence d'un seul individu atteint pour lequel nous avons identifié l'ensemble des régions homozygotes de son génome, puis recherché dans ces régions les polymorphismes candidats (SNP, petites insertions et délétions, variations structurales)

compatibles avec le phénotype observé et non présents dans les séquences d'individus contrôles non-atteints d'autres races.

Article 1 : Identification d'une délétion homozygote des exons 17 à 23 du gène de *l'integrin beta 4* chez un veau de race charolaise atteint d'épidermolyse bulleuse jonctionnelle.

Michot P, Fantini O, Braque R, Allais-Bonnet A, Saintilan R, Grohs C, Barbieri J, Genestout L, Danchin-Burge C, Gourreau JM, Boichard D, Pin D and Capitan A.

Whole-genome sequencing identifies a homozygous deletion encompassing exons 17 to 23 of the *integrin beta 4* gene in a Charolais calf with junctional epidermolysis bullosa

Michot *et al.*, Genetics Selection Evolution (2015) 47:37

Michot et al. Genetics Selection Evolution (2015) 47:37
DOI 10.1186/s12711-015-0110-z

GSE Genetics Selection Evolution

SHORT COMMUNICATION Open Access

Whole-genome sequencing identifies a homozygous deletion encompassing exons 17 to 23 of the *integrin beta 4* gene in a Charolais calf with junctional epidermolysis bullosa

Pauline Michot[1,2], Oscar Fantini[3], Régis Braque[4], Aurélie Allais-Bonnet[2,5], Romain Saintilan[1,2], Cécile Grohs[1], Johanna Barbieri[6], Lucie Genestout[7], Coralie Danchin-Burge[8], Jean-Marie Gourreau[9], Didier Boichard[1], Didier Pin[3*] and Aurélien Capitan[1,2*]

Abstract

Background: Since 2010, four Charolais calves with a congenital mechanobullous skin disorder that were born in the same herd from consanguineous matings were reported to us. Clinical and histopathological examination revealed lesions that are compatible with junctional epidermolysis bullosa (JEB).

Results: Fifty-four extended regions of homozygosity (>1 Mb) were identified after analysing the whole-genome sequencing (WGS) data from the only case available for DNA sampling at the beginning of the study. Filtering of variants located in these regions for (i) homozygous polymorphisms observed in the WGS data from eight healthy Charolais animals and (ii) homozygous or heterozygous polymorphisms found in the genomes of 234 animals from different breeds did not reveal any deleterious candidate SNPs (single nucleotide polymorphisms) or small indels. Subsequent screening for structural variants in candidate genes located in the same regions identified a homozygous deletion that includes exons 17 to 23 of the *integrin beta 4* (*ITGB4*), a gene that was previously associated with the same defect in humans. Genotyping of a second case and of six parents of affected calves (two sires and four dams) revealed a perfect association between this mutation and the assumed genotypes of the individuals. Mining of Illumina BovineSNP50 Beadchip genotyping data from 6870 Charolais cattle detected only 44 heterozygous animals for a 5.6-Mb haplotype around *ITGB4* that was shared with the carriers of the mutation. Interestingly, none of the 16 animals genotyped for the deletion carried the mutation, which suggests a rather recent origin for the mutation.

Conclusions: In conclusion, we successfully identified the causative mutation for a very rare autosomal recessive mutation with only one case by exploiting the most recent DNA sequencing technologies.

Findings

Hereditary junctional epidermolysis bullosa (JEB) is a recessive inherited blistering disorder of the skin and mucous membrane in which tissue separation occurs within the lamina lucida (i.e. under the basal plasma membrane of the basal keratinocytes and above the basement membrane) of the basement membrane zone (BMZ) at the dermal-epidermal junction [1]. This rare mechanobullous disease was previously reported to be associated with mutations in genes encoding components of the hemidesmosome anchoring complex (*ITGA6*, *ITGB4*, *COL17A1*, and *LAMA3*, *LAMB3*, and *LAMC2*, encoding the subunit polypeptides of laminin 5) in humans and in several other mammalian species [2-8]. Congenital JEB has been sporadically observed over the past 30 years in Charolais cattle but, to date, the causative mutation has not been identified.

Since 2010, four cases (three males and one female) that were born in the same French herd from consanguineous

* Correspondence: didier.pin@vetagro-sup.fr; aurelien.capitan@jouy.inra.fr
[3]Université de Lyon, VetAgro Sup, UPSP 2011-03-101 Interactions Cellules Environnement, 1 avenue Bourgelat, Marcy l'Etoile F-69280, France
[1]INRA, UMR1313 Génétique Animale et Biologie Intégrative, domaine de Vilvert, Jouy-en-Josas F-78352, France
Full list of author information is available at the end of the article

matings were reported to us. Analysis of the pedigree data revealed that all JEB-affected animals trace back, on both the maternal and paternal sides, to a single founder bull born in Great Britain in 1986 (Figure 1), which suggests an autosomal recessive mode of inheritance as the most parsimonious model.

At birth, acral clinical lesions were observed with dysungulation of the four hooves and erosions and ulcers of the skin from the carpal and tarsal joints, fetlocks, ears, eyes, and muzzle and oral cavity (i.e. nares, tongue, buccal and labial sides of the mucosa and palate) (Figure 2). In addition, three of the four cases showed ear deformities (atrophied pinna and closed ears) and one displayed major epidermal loss on the back after being licked by its mother. General signs included anorexia, apathy, emaciation and marked cutaneous pain that justified rapid euthanasia.

Histopathological examination of two cases revealed sub-epithelial splitting and blistering without keratinocyte cytolysis (Figure 3). The basal keratinocytes appeared to be intact. Periodic acid Schiff (PAS) staining was weakly positive for the basement membrane that was located at the base of the blisters. Cleft formation was sometimes present around hair follicles. Dermal inflammatory infiltrate was of varying degrees, very mild in non-ulcerated areas but marked in ulcerated areas.

Because of the rare occurrence of JEB in Charolais cattle and of rapid euthanasia of affected animals, only one case (out of the three that were born and reported at the beginning of the study) was available for DNA sampling, thus preventing the use of a classical autozygosity mapping approach [9,10]. As a consequence, we decided to sequence the whole genome (WGS) of this animal and applied an alternative strategy as described in Lupski et al. [11]. DNA was extracted from the pinna using the DNeasy Blood and Tissue Kit (Qiagen). One paired-end library with a 450-bp insert size was generated with the NEXTflex PCR-Free DNA Sequencing Kit (Bioscientific) and sequenced on one lane of the HiSeq 2000 platform (Illumina) with Illumina TruSeq V3 Kit (200 cycles).This has been submitted to the NCBI Sequence Read Archive under the accession number SRP055078. The 101-bp reads were mapped on the UMD3.1 bovine sequence assembly using BWA [12]. Reads with multiple alignments were removed (yielding a final average sequence coverage of 11.4 X) and variants were called using SAMtools [13].

Then, the genome of this consanguineous animal was screened for extended regions of homozygosity. To avoid artifactual heterozygous genotypes in homozygous regions due to pseudo-SNPs or sequencing errors, only 706 791 SNPs from the Illumina Bovine HD Beadchip were used. A total of 54 blocks with a minimal size of 1 Mb and

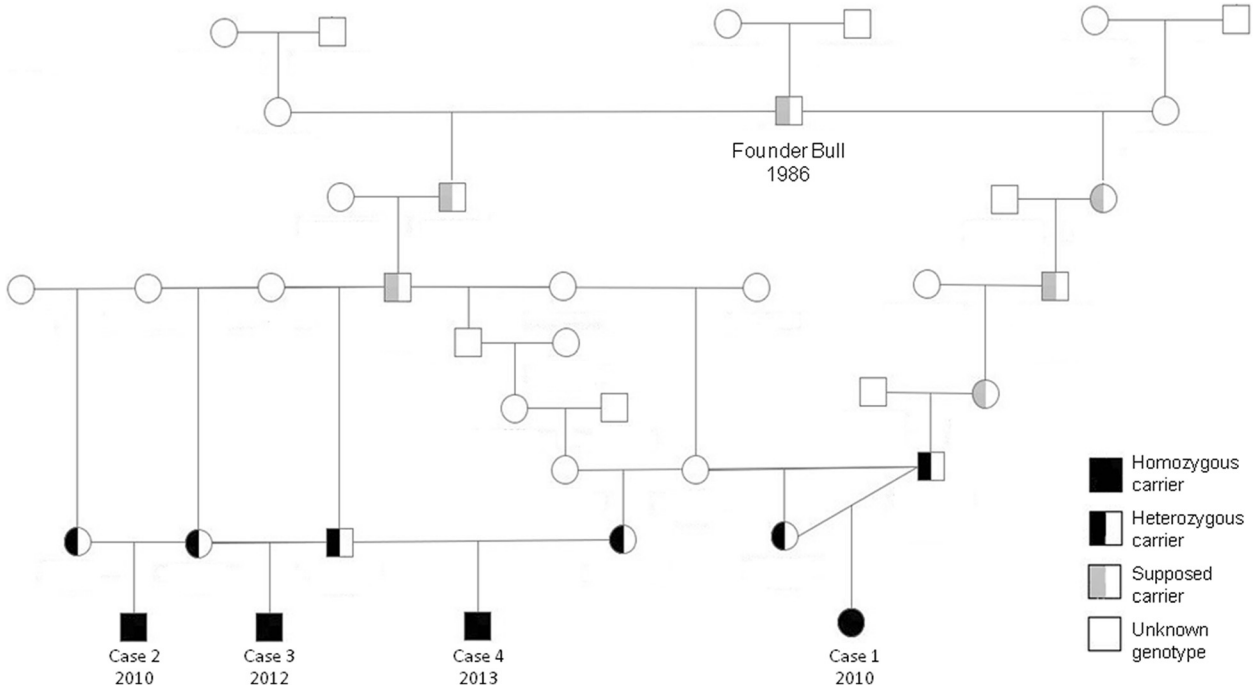

Figure 1 Pedigree of Charolais cattle with JEB. Note the high degree of consanguinity between the parents of male and female cases, which suggests an autosomal recessive mode of inheritance.

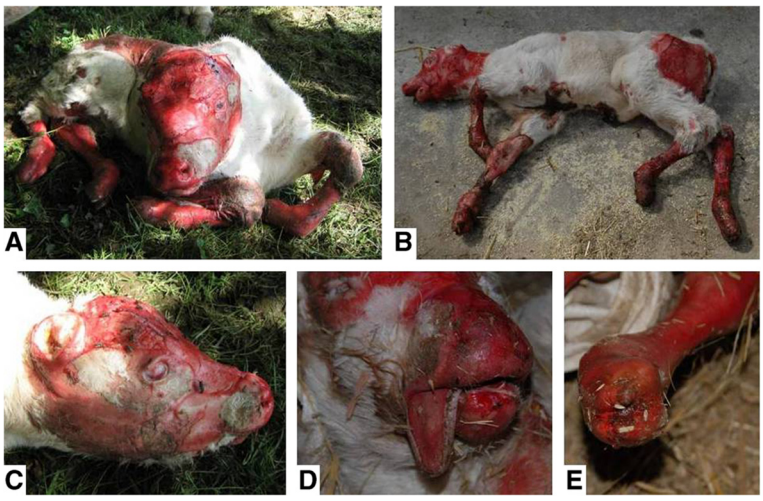

Figure 2 Clinical features of recessive JEB in Charolais cattle. (A) and **(B)** Global views of cases 3 and 2, respectively. **(C)** Head from case 3 showing atrophied pinna and skin lesions on the eyes, face and muzzle. **(D)** Lesions of the muzzle and tongue from case 1. **(E)** Forelimb from case 1 with dysungulation. These photos are personal photographs.

containing only homozygous genotypes were identified (Figure 4A).

SNPs and small Indels located in these blocks and with a quality score greater than 30 were annotated using Ensembl VEP [14] and filtered for (i) homozygous polymorphisms observed in the WGS data from eight healthy Charolais animals and (ii) homozygous or heterozygous polymorphisms found in the genomes of 234 animals from different breeds [15] (assuming that the causative mutation is recessive and specific to the Charolais cattle). No homozygous

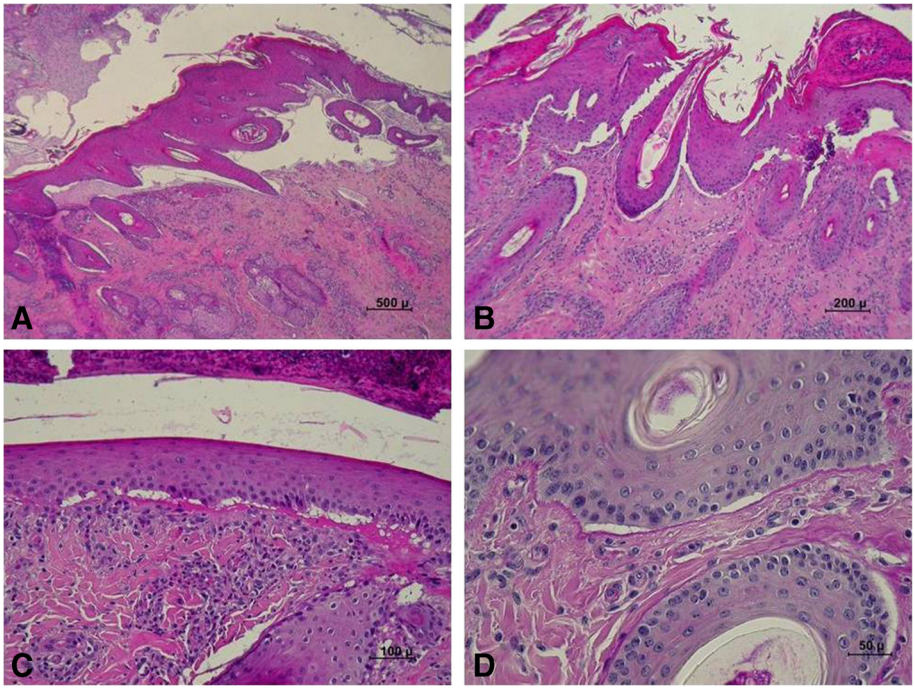

Figure 3 Histopathological features of JEB in Charolais cattle. (A) Large area of sub-epithelial splitting and blistering. **(B)** Sub-epithelial splitting and blistering, and cleft formation around hair follicles. **(C)** Vacuolation beneath basal keratinocytes of the epidermis and fibrin deposition on the dermal side of the vacuoles (5-µm section of tissue embedded in paraffin and stained with haematoxylin and eosin). **(D)** Vacuolation beneath basal keratinocytes of the epidermis and above the periodic acid Schiff (PAS)-positive basement membrane; 5-µm section of tissue embedded in paraffin and stained with PAS.

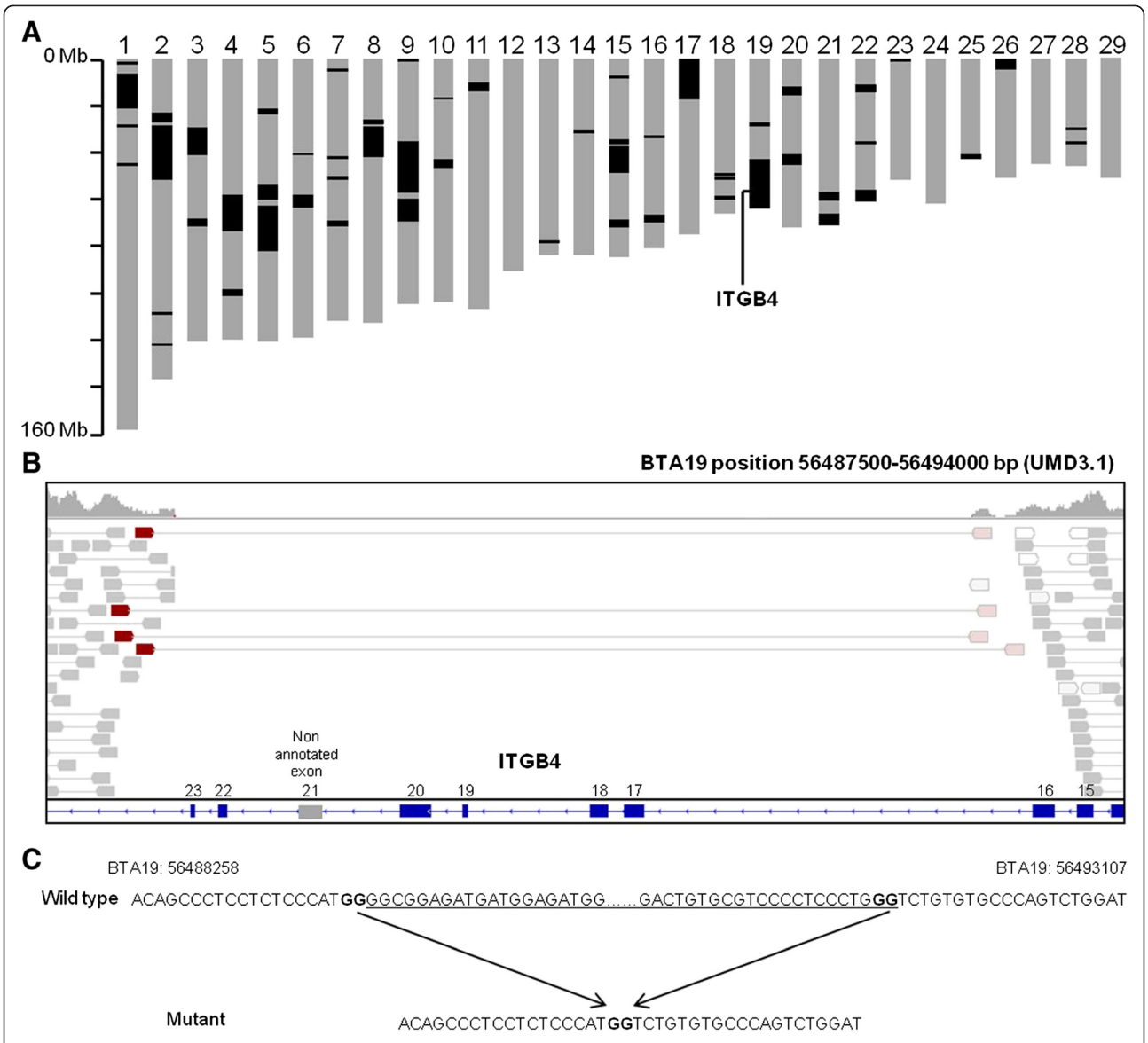

Figure 4 Identification of the mutation responsible for JEB in Charolais cattle. (A) Genome-wide screening for extended regions of homozygosity in one JEB-affected calf. Blocks of size ≥ 1 Mb are represented in black on each chromosome. **(B)** IGV snapshot showing the homozygous deletion that includes exons 17 to 23 of *ITGB4* in a JEB affected animal (modified based on results from the chromosome walking approach). **(C)** Details on the wild-type and mutant nucleotide sequences at the border of the deletion. The deleted segment is underlined; a GG doublet located on both sides of the deletion is indicated in bold.

deleterious mutation (frame-shifts, in-frame insertions or deletions, stop gain or loss of variants as well as polymorphisms that affect splice donor or splice acceptor sites and missense polymorphisms predicted to be deleterious) was found with this approach.

In a second attempt, we investigated the content of the homozygous blocks and found only one gene that was previously reported to be involved in JEB: *ITGB4*, which encodes the integrin beta 4 protein. A subsequent screening for structural variants using the Integrative Genomics Viewer (IGV) [16] enabled us to identify a 4.8-kb deletion on bovine chromosome 19 or BTA19 (g.56488278_56493087del on the UMD3.1 assembly). Since there were two gaps, and at least one artifactual segmental duplication in the current bovine assembly within the region deleted in JEB, we built a local assembly to determine the exact nature of the mutation, which finally consists in a 2831-bp deletion encompassing exons 17 to 23 of the *ITGB4* gene (Figure 4B, C and Additional file 1). This mutation was further confirmed by PCR-amplification of a 932-bp fragment that spans the deletion with the LEFT (TTCCCTGGGGGATCTGGGA) and RIGHT (CGTCTG

CGAGATCAACTACT) primers using the Go-Taq Flexi DNA Polymerase (Promega) followed by Sanger sequencing (Eurofins MWG, Ebersberg Germany).

ITGB4 encodes the beta subunit of the alpha 6 beta 4 integrin heterodimer. This transmembrane receptor is a key component of the hemidesmosome anchoring complex which connects basal keratinocytes to the basement membrane by linking the extracellular N-termini of the α and β subunits to laminin 5 whereas the intracellular C-termini are attached to the cytokeratin network via plectin or via the type XVII collagen and the bullous pemphigoid antigen 1 (BP230) [17-21] (Figure 5A). In humans, numerous mutations in *ITGB4* have been reported to cause JEB with truncating mutations being associated with a more severe (and mostly lethal) phenotype [22]. In Charolais cattle, the deletion of exons 17 to 23 of *ITGB4* is predicted to result in a complete deletion of the transmembrane domain of the protein and the joining of exon 16 to exon 24 causes a frameshift leading to the production of a protein in which all the intracellular domains in addition to the transmembrane domain (*ITGB4* p.A665Gfs*11) are missing (Figure 5B). These extensive protein modifications are assumed to impair the association of *ITGB4* with *ITGA6* and other key proteins of the hemidesmosome anchoring complex. This is consistent with the absence of hybridization signals observed by Guaguere et al. [23] after immunohistochemical characterization of the skin of a JEB-affected Charolais calf with an antibody directed towards the alpha 6 beta 4 integrin complex.

Genotyping of a second case available for DNA sampling (born at the end of our study) and of the

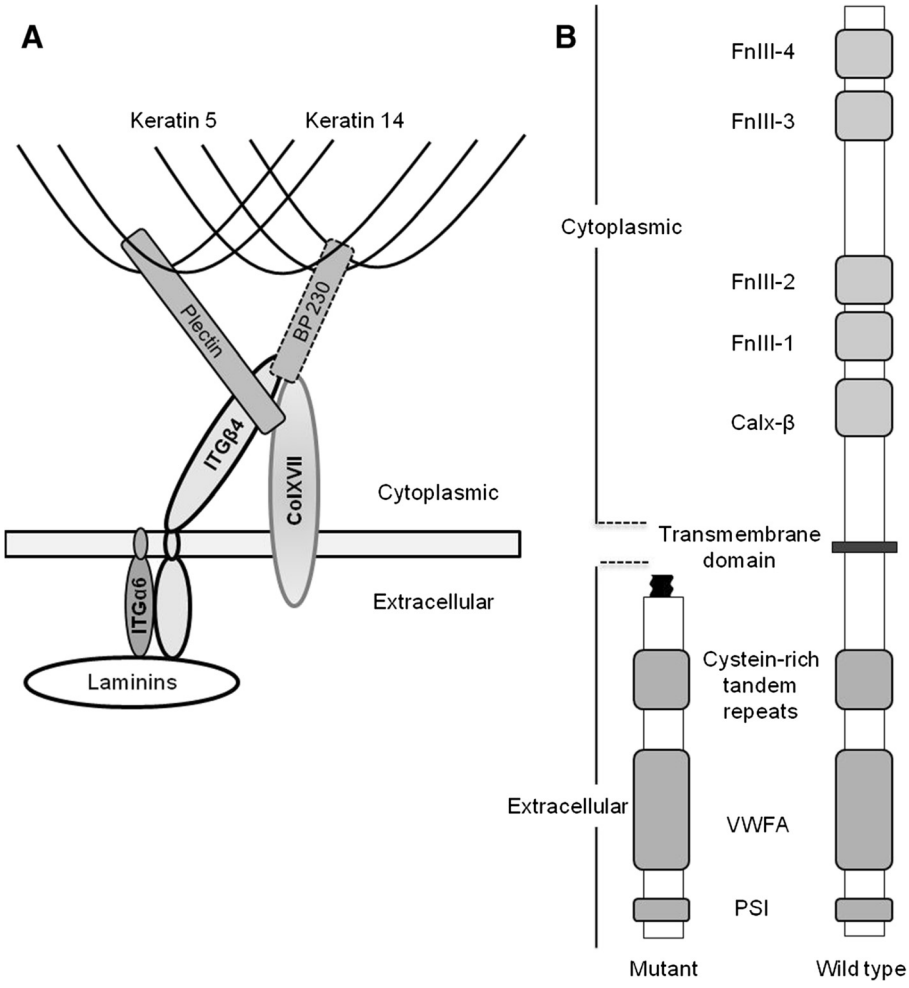

Figure 5 Schematic representation of the components of the hemidesmosomes and of the *ITGB4* protein. (A) Schematic representation of the components of the hemidesmosomes (adapted from http://xtal.cicancer.org/research.html). **(B)** Comparison between the structure of the mutant (predicted) and wild-type integrin β4 subunit (http://www.uniprot.org/uniprot/P16144). The wild-type integrin β4 comprises: (i) an extracellular region that contains the N-terminal plexin-semaphorin-integrin (PSI) and von Willebrand factor type A (VWFA) domains as well as a cysteine-rich repeat region; (ii) a transmembrane domain and (iii) a cytoplasmic region that includes a calx-beta (calx- β) domain and four fibronectin III-like domains (FnIII-1 to 4). The transmembrane domain and cytoplasmic region are predicted to be totally absent in the mutated β4 subunit.

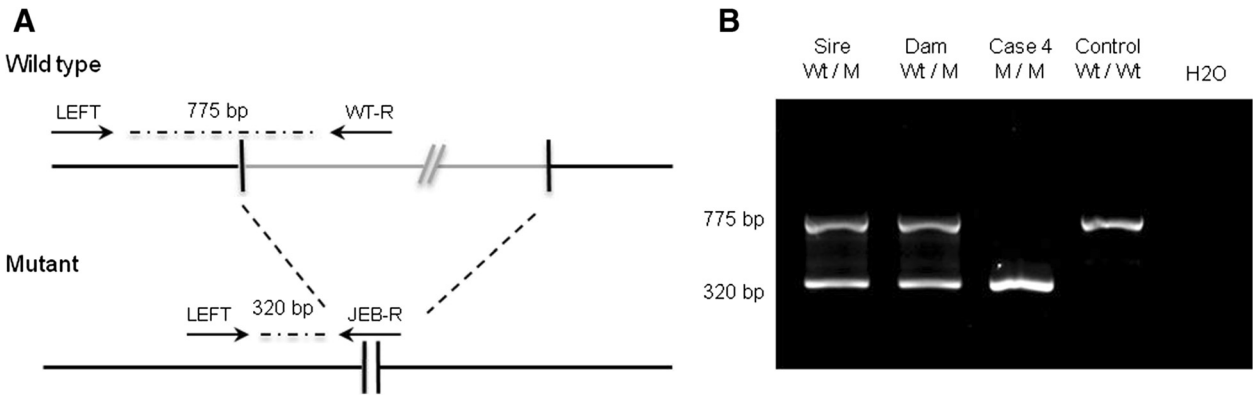

Figure 6 Genotyping of the deletion using a 3-primer PCR system. (A) Details on the design of the 3-primer PCR system. **(B)** Genotyping data of case 4, its parents, and a homozygous wild-type animal (based on haplotype information).

six parents of the affected animals using a 3-primer PCR system (products amplified with primers LEFT, WT-R: TCTGCCCCACATGAATGCTT and JEB-R: AGACTGGG CACACAGACCAT using the same polymerase and revealed by electrophoresis on an ethidium-bromide-stained 2% agarose gel) revealed a perfect association between this mutation and the assumed genotypes of the individuals (Figure 6A and B).

Taken together, these arguments strongly support that this deletion within *ITGB4* is responsible for autosomal recessive JEB in Charolais cattle.

In an attempt to estimate the frequency of this mutation in the Charolais population that is bred by artificial insemination (AI), four parents were genotyped with the Illumina BovineSNP50 Beachip and phased as described in Boichard et al. [24] together with 6870 Charolais animals previously genotyped with the same array for genomic selection. Analysis of these data identified 44 animals that were all heterozygous for a rare (frequency = 3.2 %) 5.6-Mb segment (105 markers between positions 51 796 076 and 57 397 180 Mb on BTA19) that was shared identical by descent (IBD) with the haplotype carrying the *ITGB4* deletion. Surprisingly, subsequent genotyping of 16 of these animals with our 3-primer PCR system revealed that none were carriers of the deletion suggesting a rather recent origin for this mutation on the scale of the Charolais breed's history. While our analysis also suggests the absence of the *ITGB4* deletion in the French AI population, it would be relevant to genotype the French natural mating population. Indeed the founder bull descends from animals exported to Great Britain at the end of the 1960's from this population and it has been used as bull sire in French breeding herds in the 1990's.

In conclusion, with only one case, we identified a mutation which appears to be necessary and sufficient to cause autosomal recessive junctional epidermolysis bullosa by exploiting the most recent DNA sequencing technologies. Targeted genotyping of at risk pedigrees with the genetic test that we developed will allow the rapid eradication of this rare genetic disease in Charolais cattle.

Additional file

Additional file 1: Building of a local sequence assembly for the region deleted in the JEB calf. This additional file provides details on the methodology used to build the local assembly of the deleted region in the JEB calf and the nucleotide sequence associated.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AC and DP conceived and coordinated the study. RB performed clinical examinations. DP and OF performed histological analyses. PM, AC, AB, RS, CG, JB and DB participated in the genetic study. RB, JMG, LG, CD and DB contributed material and data. AC, PM and DP wrote the manuscript and DB critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The authors are grateful to the breeders, C Bodet and M Mayet, for excellent collaboration. We would also like to thank the Charolais Univers and Gènes Diffusion breeding companies for giving us access to DNA samples and genotyping data generated for genomic selection purposes and Y Amigues for retrieving these DNA samples from the LABOGENA DNA bank. Finally, the help of S Jurado (chambre d'Agriculture de la Nièvre) and G Fosserz (Herd Book Charolais) for reconstituting the pedigree of the affected family is highly appreciated. P Michot is recipient of a PhD grant from ALLICE and Apis Gène.

Author details
[1]INRA, UMR1313 Génétique Animale et Biologie Intégrative, domaine de Vilvert, Jouy-en-Josas F-78352, France. [2]ALLICE, 149 rue de Bercy, Paris F-75012, France. [3]Université de Lyon, VetAgro Sup, UPSP 2011-03-101 Interactions Cellules Environnement, 1 avenue Bourgelat, Marcy l'Etoile F-69280, France. [4]Cabinet des Vignes de la Fontaine, 41 rue du faubourg de Moulins, Saint-Pierre le Moutier F-58240, France. [5]UMR 1198 Biologie du Développement et Reproduction, domaine de Vilvert, Institut National de la Recherche Agronomique, Jouy-en-Josas F-78352, France. [6]INRA, UMR1388 GenPhySE, GeT-PlaGe, Castanet-Tolosan F-31320, France. [7]LABOGENA DNA, domaine de Vilvert, Jouy-en-Josas F-78352, France. [8]Institut de l'Elevage, 149 rue de Bercy, Paris 12 F-75595, France. [9]Unité de Pathologie du Bétail, Ecole Nationale Vétérinaire d'Alfort, 7 avenue du Général de Gaulle, Maisons-Alfort F-94704, France.

Received: 5 November 2014 Accepted: 3 March 2015
Published online: 03 May 2015

References

1. Fine JD, Eady RA, Bauer EA, Bauer JW, Bruckner-Tuderman L, Heagerty A, et al. The classification of inherited epidermolysis bullosa (EB): report of the third international consensus meeting on diagnosis and classification of EB. J Am Acad Dermatol. 2008;58:931–50.
2. Pulkkinen L, Christiano AM, Airenne T, Haakana H, Tryggvason K, Uitto J. Mutations in the gamma-2 chain gene (*LAMC2*) of kalinin/laminin 5 in the junctional forms of epidermolysis bullosa. Nat Genet. 1994;6:293–8.
3. McGrath JA, Pulkkinen L, Christiano AM, Leigh IM, Eady RAJ, Uitto J. Altered laminin 5 expression due to mutations in the gene encoding the beta-3 chain (*LAMB3*) in generalized atrophic benign epidermolysis bullosa. J Invest Derm. 1995;104:467–74.
4. Vidal F, Aberdam D, Miquel C, Christiano AM, Pulkkinen L, Uitto J, et al. Integrin beta-4 mutations associated with junctional epidermolysis bullosa with pyloric atresia. Nat Genet. 1995;10:229–34.
5. McGrath JA, Gatalica B, Christiano AM, Li K, Owaribe K, McMillan JR, et al. Mutations in the 180-kD bullous pemphigoid antigen (BPAG2), a hemidesmosomal transmembrane collagen (COL17A1), in generalized atrophic benign epidermolysis bullosa. Nat Genet. 1995;11:83–6.
6. Vidal F, Baudoin C, Miquel C, Galliano M-F, Christiano AM, Uitto J, et al. Cloning of the laminin alpha-3 chain gene (*LAMA3*) and identification of a homozygous deletion in a patient with Herlitz junctional epidermolysis bullosa. Genomics. 1995;30:273–80.
7. Ruzzi L, Gagnoux-Palacios L, Pinola M, Belli S, Meneguzzi G, D'Alessio M, et al. Homozygous mutation in the integrin alpha-6 gene in junctional epidermolysis bullosa with pyloric atresia. J Clin Invest. 1997;99:2826–31.
8. Bruckner-Tuderman L, McGrath JA, Clare Robinson E, Uitto J. Animal models of epidermolysis bullosa: update 2010. J Invest Derm. 2010;130:1485–8.
9. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987;236:1567–70.
10. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, et al. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet. 2008;40:449–54.
11. Lupski JR, Reid JF, Gonzaga-Jauregui C, Rio Deiros D, Chen DCY, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. N Engl J Med. 2010;362:1181–91.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 1000 genome project data processing subgroup: the sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
14. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. Bioinformatics. 2010;26:2069–70.
15. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.
16. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
17. Hopkinson SB, Findlay K, DeHart GW, Jones JC. Interaction of BP180 (type XVII collagen) and alpha6 integrin is necessary for stabilization of hemidesmosome structure. J Invest Dermatol. 1998;111:1015–22.
18. Hopkinson SB, Jones JC. The N terminus of the transmembrane protein BP180 interacts with the N-terminal domain of BP230, thereby mediating keratin cytoskeleton anchorage to the cell surface at the site of the hemidesmosome. Mol Biol Cell. 2000;11:277–86.
19. Fontao L, Favre B, Riou S, Geerts D, Jaunin F, Saurat JH, et al. Interaction of the bullous pemphigoid antigen 1 (BP230) and desmoplakin with intermediate filaments is mediated by distinct sequences within their COOH terminus. Mol Biol Cell. 2003;14:1978–92.
20. Koster J, Geerts D, Favre B, Borradori L, Sonnenberg A. Analysis of the interactions between BP180, BP230, plectin and the integrin alpha6beta4 important for hemidesmosome assembly. J Cell Sci. 2003;116:387–99.
21. De Pereda JM, Lillo MP, Sonnenberg A. Structural basis of the interaction between integrin alpha6beta4 and plectin at the hemidesmosomes. EMBO J. 2009;28:1180–90.
22. Pulkkinen L, Rouan F, Bruckner-Tuderman L, Wallerstein R, Garzon M, Brown T, et al. Novel *ITGB4* mutations in lethal and nonlethal variants of epidermolysis bullosa with pyloric atresia: missense versus nonsense. Am J Hum Genet. 1998;63:1376–87.
23. Guaguere E, Berg K, Degorce-Rubiales F, Spadafora A, Meneguzzi G. Junctional epidermolysis bullosa in a Charolais calf with deficient expression of integrin α6ß4. Vet Dermatol. 2004;15:28.
24. Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, et al. Genomic selection in French dairy cattle. Anim Prod Sci. 2012;52:115–20.

Dans ce premier article, nous avons mis en place une stratégie alternative à la cartographie classique basée sur l'analyse directe des données de séquence. Celle-ci a permis l'identification de la mutation causale d'une anomalie récessive rare à partir d'un seul génome d'individu atteint séquencé. Coïncidence surprenante, cette délétion dans le gène *ITGB4* a été identifiée indépendamment de notre étude par Peters *et al.* (2015) à partir d'un seul cas de veau atteint d'EBJ et en appliquant une stratégie similaire. Ce type de stratégie basée sur les données WGS, est appliquée couramment ces dernières années en génétique humaine où les anomalies génétiques sont souvent restreintes à un ou deux atteints dans une même famille. En effet, l'accumulation des données de séquence au travers de différents projets comme le « 1000 génomes humains » permet l'accès à une base d'individus contrôles suffisamment grande pour réduire le nombre de variants candidats identifiés par séquençage de l'exome ou du génome complet d'un individu atteint (Ng *et al.*, 2010).

Comme le montre l'étude de l'EBJ, il est aujourd'hui envisageable d'identifier une anomalie génétique sans forcément passer par une étape préliminaire de cartographie avec marqueurs. Avec la diminution des coûts de re-séquençage, l'augmentation des données de génome complet qui en résulte et le partage des données au sein de consortiums (« 1000 bull genomes », Daetwyler *et al.*, 2014) elles offrent de nouvelles voies pour l'identification d'anomalies génétiques rares chez le bovin.

CHAPITRE 4 : IDENTIFICATION D'UNE MUTATION CANDIDATE POUR UN PHENOTYPE DE MORTALITE EMBRYONNAIRE EN RACE MONTBELIARDE

Avec le développement de la sélection génomique et du génotypage, les recherches d'haplotypes avec déficit significatif en homozygotes, dont le principe a été présenté plus haut (voir Chapitre 1, III-B-2), se sont révélées efficaces dans la détection des anomalies génétiques récessives. De plus elles sont complémentaires des approches précédentes, en permettant d'identifier des anomalies passant inaperçues pour les observatoires, en particulier les mortalités de l'embryon et du fœtus au cours de la gestation, qui impactent fortement la fertilité des troupeaux.

En France, Fritz *et al*. (2013) ont réalisé une détection de ce type dans les races laitières françaises Holstein, Normande et Montbéliarde. Pour la race Montbéliarde, deux principaux QTL de mortalité embryonnaire ont été révélés : MH1 sur le chromosome 19 et MH2 sur le chromosome 29. L'analyse des données de séquence des taureaux porteurs disponibles à l'époque avait alors permis de proposer deux mutations non-sens affectant les gènes codant pour la *Sex-hormone binding globulin* (SHBG) et le transporteur *Solute carrier Family 37 A2* (SLC37A2), en tant que mutations candidates. Ces deux mutations, non confirmées en 2013, avaient été intégrées à la puce de génotypage Illumina EuroG10K pour assurer un suivi dans la population. Parmi les animaux génotypés, des individus vivants homozygotes pour la mutation du gène SHBG ont été observés. Ceci nous a conduits à invalider ce candidat et à poursuivre l'étude de l'haplotype MH1.

L'article qui constitue ce quatrième chapitre de ma thèse présente les travaux complémentaires réalisés dans ce cadre. En s'appuyant sur les données de génotypage et le suivi des haplotypes recombinant avec MH1, nous avons réduit l'intervalle de localisation de la mutation et nous proposons comme nouvelle mutation causale une substitution d'un acide aminé dans la Phosphoribosylformylglycinamidine synthase (PFAS), une protéine qui intervient dans la voie de bio-synthèse des purines. Dans cette étude, la combinaison des données de génotypage et de séquençage, nous a permis d'identifier la mutation candidate et de fournir un test efficace pour la contre sélection de cette anomalie létale et fréquente (7%) en race Montbéliarde.

Article 2 : Une substitution dans le gène *PFAS* est probablement responsable de la mortalité embryonnaire associée à l'haplotype MH1 en race Montbéliarde

Michot P.[1,2], Fritz S. [1,2], Barbat A. [1], Boussaha M.[1] , Deloche MC.[2], Grohs C. [1], Hoze C.[1,2],Le Berre L[2,3]., Le Bourhis D.[2,3], Desnoes O.[2 ,3], Salvetti P.[2,3], Schibler L.[2], Boichard D. [1] and Capitan A. [1,2]

A missense mutation in *PFAS* is likely causal for embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle. Journal of Dairy Science, submitted in January 2017.

Interpretive Summary

A missense mutation in *PFAS* is likely causal for embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle. *By Michot et al., page xxx.*

MH1 is a haplotype on chromosome 19 associated with reduced fertility and a deficit of homozygotes in Montbéliarde cattle. It was initially attributed to a nonsense mutation in SHBG gene. Here, we demonstrate that the causative mutation is not this *SHBG* mutation but more likely a missense mutation affecting a conserved amino-acid of PFAS, a protein involved in the *de novo* biosynthesis of purines. We also show that embryonic mortality occurs between 7 and 35 days. Because of its high frequency (7%), selection against this deleterious allele is highly recommended and under way.

A missense mutation in *PFAS* is likely causal for embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle

Pauline Michot,*† Sébastien Fritz,*† Anne Barbat,* Mekki Boussaha,* Marie-Christine Deloche,*† Cécile Grohs,* Chris Hoze,*† Laurène Le Berre,†‡ Daniel Le Bourhis,†‡ Olivier Desnoes,†‡ Pascal Salvetti,†‡ Laurent Schibler,† Didier Boichard,* and Aurélien Capitan*†[1]

*****UMR GABI, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy en Josas, France

†Allice, 75595 Paris, France

‡Allice, Station de phénotypage, 37380 Nouzilly, France

[1]Corresponding author: aurelien.capitan@inra.fr

Aurélien Capitan, Bat 211, INRA-CRJ, Domaine de Vilvert, 78352 Jouy-en-Josas cedex, France

+33 1 34 65 26 45

ABSTRACT

A candidate mutation in the *sex hormone binding globulin* (*SHBG)* gene was proposed in 2013 to be responsible for the MH1 recessive embryonic lethal locus segregating in the Montbeliarde breed. In this follow-up study, we exclude this candidate variant since healthy homozygous for the derived allele were observed in large scale genotyping data generated in the framework of the genomic selection program. We fine map the MH1 locus in a 702-kb interval and analyze genome sequence data from the 1000 bull genomes project and 54 Montbeliard bulls (including 14 carriers and 40 non carriers). We report the identification of a strong candidate mutation in the gene encoding the *Phosphoribosylformylglycinamidine synthase* (PFAS), a protein involved in *de novo* purine synthesis. This results in the substitution of an arginin residue that is entirely conserved among eukaryotes by a cystein (p.R1205C), and located in a Class I glutamine amidotransferase-like domain. No homozygote for the derived allele was observed in a large population of more than 20,000 individuals, in spite of a 7.2% allelic frequency. Genotyping of 18 embryos born from heterozygous parents and analysis on non-return rates suggested that most of homozygous carriers died between 7 and 35 days post-insemination. The identification of this strong candidate mutation will enable the accurate testing of the reproducers and the efficient selection against this recessive embryonic lethal defect in the Montbeliarde breed.

Key words: embryonic lethality, deficit in homozygote, whole genome sequencing, large scale genotyping, PFAS

INTRODUCTION

The recent development of genomic selection in cattle has generated an unprecedented large scale genotyping activity, thus opening new possibilities in the search of causal variants. In 2011, VanRaden *et al.* proposed to use these genotyping data to identify haplotypes displaying a significant deficit in homozygous animals and thus to map putative loci responsible for recessive embryonic or perinatal lethality. A number of similar studies were subsequently carried out in different breeds and subpopulations (Fritz *et al.*, 2013; Sahana *et al*., 2013; Sonstegard *et al.*, 2013; McClure *et al.*, 2014; Venhoranta *et al.*, 2014; Pausch *et al.*, 2015; Adams *et al.*, 2016; Menzi *et al.*, 2016; Schütz *et al.*, 2016; Schwarzenbacher *et al.*, 2016) leading to the identification of tens of new recessive defects. Using this strategy, we previously reported the identification of 11 haplotypes displaying a significant deficit in homozygous in the Montbeliarde breed (Fritz *et al.*, 2013). Out of these, MH1 and MH2 were the two most frequent haplotypes (9 and 7%, respectively) and showed a significant negative effect on calving rate confirming their association with embryonic lethal mutations. Then, mining whole genome sequences (WGS) from carriers and non-carrier bulls, we identified two nonsense mutations in the Sex-hormone-binding globulin (chr19 g.27956790C>T; SHBG p.Q52X) and Solute carrier Family 37 A2 protein (g. chr29:28879810C>T; SLC37A2 p.R12X) genes. We confirmed their association with the MH1 and MH2 haplotypes but did not validate their causality with large scale genotyping or functional studies. Since then both SNPs have been added to the Illumina EuroG10k SNP custom Chip, widely used for genomic selection in France, and generated genotyping data for tens of thousands of animals. Within this large population we identified 242 homozygous carriers for the *SHBG* nonsense mutation suggesting that it was not the causative mutation for MH1-linked embryonic lethality. Similar observations were also made in Simmental (Pausch, personal communication) and Vorderwälder breeds (Reinartz *et al.*, 2016).

The purpose of this follow-up study was threefold: (1) to fine map the MH1 locus, (2) to identify the causative mutation and (3) to further characterize its phenotypic effects using the large amount of phenotypes, genotypes and whole-genome sequences that have accumulated since the initial study.

MATERIALS AND METHODS

Genotyping Data

Genotypes from 115,167 Montbéliarde animals analyzed from 2008 to 2015 with different Illumina chips (BovineSNP50 Beadchip (50K) v1 and v2, LD chip, HD chip, EuroG10K) were available from the French genomic evaluation database. These were checked for quality and processed in the framework of the French genomic selection pipeline (Boichard *et al.*, 2012), including imputation and phasing with FImpute (Sargolzaei *et al.*, 2014). Then phased 50k genotype data from each animal were screened for the MH1 lethal haplotype as described by Fritz *et al.* (2013) to determine their carrier or

non-carrier status. This haplotype consists in 27 consecutive markers from Hapmap51730-BTA-44937 (position 27,643,677 bp) to ARS-BFGL-NGS-111936 (position 29,383,514 bp) on chromosome 19.

The SHBG p.Q52X (g.27956790C>T) polymorphism previously described by Fritz *et al.*, (2013) and a new candidate polymorphism for MH1 (g.28511199C>T; PFAS p.R1205C) were included in two consecutive custom designs of the Illumina EuroG10K custom SNP BeadChip. At the time of the study, genotype information for these two variants was available for 47,871 and 20,350 Montbeliarde animals, respectively.

Whole Genome Sequencing Data

This study included WGS from 274 individuals from 15 different French dairy and beef breeds (Boussaha *et al.*, 2016) and from 1022 additional animals from run4 of the 1000 bull genomes consortium (Daetwyler *et al.*, 2014). The database contained WGS from 54 Montbeliarde artificial insemination (AI) bulls, of which 14 were MH1 carriers. For each animal, paired-end libraries of 150 to 400 bp DNA fragments were prepared using Illumina TruSeq DNA Sample Prep Kit, controlled for quality and sequenced on an Illumina HiSeq 2000 sequencing platform. The 100 pb paired-end reads generated were then aligned on the UMD3.1 bovine genome reference assembly with the Burrows-Wheeler Alignment tool (BWA v0.6.1-r104) (Li and Durbin, 2009). After removing potential PCR duplicates and reads with a low mapping quality (<=30), SNPs (single nucleotide polymorphisms) and small InDels (insertions and deletions) were detected from the resulting BAM files using the Genome Analysis Tool Kit 2.4–9 (GATK) version and GATK-UnifiedGenotyper as SNP caller (McKenna *et al.*, 2010). Bioinformatics detection of large genomic variations was carried out only on the 274 individuals previously mentioned, including the 54 Montbeliarde bulls. Multi-sample variant calling was performed using Pindel software, v. 0.2.4y (Ye *et al.*, 2009) as described in Boussaha *et al.* (2015). Structural variations were subsequently visualized and validated using IGV (integrative genomics viewer, Robinson *et al.*, 2011).

Fine Mapping of the MH1 Locus Using Recombinant Haplotypes

In line with previous studies (Sonstegard *et al.*, 2013; Adams *et al.*, 2016), we screened our population for animals which are compound heterozygous for the original lethal haplotype and a recombining haplotype. Since these animals are alive, they cannot be homozygous for the lethal mutation. This leads to the exclusion from the location interval of the identical-by-descent segment for which they are homozygous. Here we took advantage of the large contribution of the MH1 carrier bull BOISLEVIN (MONFRAM00186006232) to the Montbeliarde population. We defined two regions of 40 markers denoted hapL and hapR, located 3 Mb respectively upstream and downstream of the MH1 haplotype. We subsequently identified the hapL and hapR source haplotypes of BOISLEVIN linked to MH1. Among his genotyped descendants, we retrieved all the individuals homozygous for hapL. Then the

hapL haplotype was extended toward MH1 one SNP at a time and individuals carrying a haplotype different from the initial one, therefore recombinant, were eliminated. The process was continued until no individual was retained. The same approach was repeated for hapR. On each side, the last remaining individuals defined the left or right boundary of the shortest interval carrying the lethal variant.

Annotation and Filtering of Genetic Variants

Variants from the genomes of the 14 Montbeliarde MH1 carriers were filtered to retain only those located within the critical interval and that were never observed in the homozygous state in the whole dataset. Remaining variants were annotated using Ensembl Variant Effect Predictor pipeline (VeP) on the Ensembl v81 transcript set (MacLaren *et al.*, 2010). The effect of non synonymous coding variants was predicted with the SIFT software (Kumar *et al.*, 2009). Finally, the concordance between the genotypes of the 54 Montbeliard bulls and their status for the MH1 locus was estimated for each variant using a Fisher's exact test. A Bonferroni correction was applied to account for multiple testing.

Multiple Alignment of Protein Sequences

Sequences of the bovine PFAS protein and its orthologs in 16 eukaryote species available in Ensembl (http://www.ensembl.org) were aligned using Clustal Omega software (http://www.ebi.ac.uk/Tools/msa/clustalo).

Evaluation of the Effect of the PFAS Mutation on Non-Return Rate at 56 Days

In our previous study (Fritz *et al.*, 2013), analysing artificial insemination data, we demonstrated a negative effect of the MH1 haplotype on calving rate (i.e. presence/absence of a calf after complete gestation) which was consistent with recessive embryonic lethality. Here, to precise the stage of embryonic loss, we evaluated the effect of the PFAS mutation on non-return rate at 56 days (NRR56). NRR56 was coded 0 if a second insemination was observed in less than 56 days after the first insemination and 1 otherwise. Because this locus has a recessive effect, the expected decrease in fertility in mating at risk is 0.25*µ between two carriers (with µ being the average conception rate), and $0.5*0.5*(\frac{1}{2-f})*$ µ between a carrier bull and the daughter of a carrier (where $\frac{1}{2-f}$ is the proportion of carriers among daughters of carriers bull and of dams of unknown genotype, and f the frequency of the deleterious allele). The mean conception rate µ reaches 44% in lactating cows and 55% in virgin heifers. The model used to analyze NRR56 was the same as in Fritz *et al.* (2013). It accounted for several environmental effects, the fixed effect of the combination of the PFAS status of mated bulls with that of the sire of the cow, and the random effects of the sire of the embryo and the sire of the dam.

Genotypes for the PFAS mutation were obtained directly from the Illumina EuroG10K SNP chip or indirectly through a haplotypic test based on 50k genotypes. This test was developed with a reference population of 20,350 Montbeliarde animals genotyped with the EuroG10K chip and imputed to the 50k. Only inseminations before May 2016 with known status for (i) males and females or (ii) males and sires of females were considered. In total, 6,176,305 NRR56 data were analyzed, including 270,113 from matings at risk. Tests were carried out separately for lactating cows and heifers. Finally a t-test was used to determine any significant difference in conception rate between matings at risk and control groups.

Production and Genotyping of Embryos from Mating at Risk

Four heifers heterozygous for mutation g.28511199C>T on chromosome 19 (PFAS p.R1205C) and for the MH1 haplotype were selected among the animals genotyped for genomic selection. In the experimental station of Allice (Nouzilly, France), they were superovulated using standard protocols (Munoz *et al.*, 2014) and inseminated with the semen of two different bulls of the same genotype. On day 7, embryos were collected. Their DNA was extracted and a whole genome amplification was carried out with the Repli-g® Mini Kit (Qiagen). Genotyping for the g.28511199C>T mutation was performed by PCR and Sanger sequencing. Forward (PFAS_F: AGTCCCCTTTCACTCCAGGT) and reverse (PFAS_R: TGGAGTTCCGGAGAAGAAAA) primers were designed from the UMD3.1 bovine genome assembly using Primer3 software (Rosen *et al.*, 2000). PCR amplification was performed using the Go-Taq Flexi DNA Polymerase (Promega) according to the manufacturer's instructions on a Mastercycler pro thermocycler (Eppendorf). The PCR products were subsequently purified and bidirectionally sequenced by Eurofins MWG (Germany) using conventional Sanger sequencing. Polymorphisms were detected with the novoSNP software (Weckx *et al.*, 2005).

RESULTS

Invalidation of the SHBG p.Q52X Mutation

Since 2013, 47,871 Montbeliarde animals have been genotyped for the SHBG p.Q52X polymorphism with the Illumina EuroG10K custom SNP BeadChip used for genomic selection. Comparison between the genotypes of these animals and their status for the MH1 haplotype revealed a strong but incomplete linkage. Notably, while no homozygous carrier of the MH1 haplotype was observed, we identified 242 animals homozygous for the *SHBG* nonsense mutation. This led us to consider that the SHBG p.Q52X mutation was not the causative mutation for MH1-associated embryonic lethality (Table 1).

Fine Mapping of the MH1 Causal Mutation

In line with previous studies (Sonstegard *et al.* 2013; Adams *et al.*, 2016), we screened our population for compound heterozygotes carrying the original lethal haplotype and a recombining haplotype among the inbred descendants of funder bulls who spread the mutation (e.g. BOISLEVIN in our situation; see

methods and Figure 1a). Such animals cannot be homozygous for the lethal mutation and thus provide useful information to fine map the MH1 mutation. In doing so, we identified several recombining haplotypes, among which two enabled us to reduce the critical interval to a region of 702 kb (Chr29:28,450,297-29,152,662) from marker ARS-BFGL-NGS-95155 to ARS-BFGL-NGS-43359 (Figure 1a and 1b). Notably, this interval was included inside the MH1 source haplotype defined by Fritz *et al*. (2013) and excluded the *SHBG* mutation. These two recombining haplotypes, mostly diffused by the AI sires REDON (MONFRAM002529434146, for HapR1) and RAPALLO (MONFRAM007120640289, for HapR2), were observed in the compound heterozygous state with MH1 (i.e. MH1/HapR1 and MH1/HapR2) in 259 and seven individuals, respectively. Of note, MH1/HapR1, and to a lesser extent HapR1/HapR1, were the two most frequent genotypes observed among the homozygotes for the *SHBG* nonsense mutation mentioned previously. Finally, by analyzing pedigree and haplotype information, we were able to trace back each recombination event and to confirm that they were posterior to BOISLEVIN (Figure 1c).

Screening of WGS Data Identifies a Strong Candidate Polymorphism in the PFAS Gene

In this step, we considered only variants from 14 MH1 carriers that were never found in the homozygous state in 40 non carrier Montbeliarde bulls and in additional controls. SNP and indel information was available for 1242 genomes, whereas structural variations were identified only in 220 genomes (see methods). In the critical interval, a total of 27 variants displayed a strong association with the MH1 haplotype (Fisher exact test p-values between 6.46×10^{-8} and 1.86×10^{-10} after Bonferroni correction; Figure 2a and Supplementary table 1). These comprised 25 small noncoding variants located in regions poorly conserved among mammals and two missense mutations. Among them, mutation g.28979780C>T (PIK3R5 *p.*R363H) located in the phosphoinositide-3-kinase regulatory subunit 5 gene was predicted to be tolerated according to SIFT (Kumar *et al.*, 2009). Indeed, the alternative allele is not conserved in mammals and corresponds to the ancestral allele in humans and other species. In contrast, mutation g.28511199C>T is predicted to be damaging and to result in the substitution of an arginin which is entirely conserved among eukaryotes by a cystein in a Class I glutamine amidotransferase-like domain of the phosphoribosylformylglycinamidine synthase (PFAS p.R1205C). According to raw WGS data, this mutation showed nearly but not complete association with the MH1 haplotype. Indeed, one MH1 carrier bull displayed low local sequencing coverage (5 x) and was genotyped as homozygous for the ancestral allele whereas a second carrier, BOISLEVIN, had missing genotype (0x). Subsequent genotyping of these two bulls by PCR-Sanger sequencing actually revealed that they were both heterozygous for the *PFAS* g.28511199C>T variant, thus demonstrating a perfect association between the mutation and the MH1 haplotype.

Large Scale Genotyping Reveals Complete Linkage between PFAS g.28511199T Allele and the Embryonic Lethal Mutation

For further verification, we genotyped the *PFAS* g.28511199C>T polymorphism in a large population of 20,350 Montbeliarde cattle using the Illumina EuroG10K BeadChip. The frequency of the g.28511199C>T was 7.2%. No homozygous mutant was observed whereas 105 were expected in this dataset (χ^2 p-value=2.51×10^{-28}; Table 2). This demonstrates that *PFAS* g.28511199T allele is in complete linkage disequilibrium with the embryonic lethal mutation.

Analysis of Non Return Rate and Genotyping of Embryos Suggests that Mortality Occurs between 7 and 35 Days of Development

In 2013, we showed a negative effect of MH1 haplotype on calving rate in mating at risk which was consistent with recessive embryonic lethality. This observation is also consistent with the complete absence of homozygous animals for the candidate mutation in *PFAS* gene. To narrow down the stage at which occurs the death of homozygous embryos, we tested the effect of the g.28511199T allele on the non return rate at 56 days after insemination (NRR56; Table 3). In comparison with the control group, we observed a significant loss in NRR56 for both heifers and adult cows in mating at risk between *PFAS* carriers as well as between *PFAS* carriers and daughters of *PFAS* carriers. Estimated values were close to the expected effect under the assumption of complete lethality in homozygous embryos (-12,5% and -6.5% respectively with an average conception rate of 50%). Considering that the reproductive cycle is on average of 21 days in females, this result indicates that embryo development stopped in the first 35 days of gestation. For further verification, we produced embryos from mating at risk between heterozygous carriers of the PFAS g.28511199C>T polymorphism. In total, 18 embryos were recovered at seven days of gestation and genotyped (Table 4). At this stage, we observed four homozygous for the derived allele, a proportion in agreement with Mendel's rules. None of them displayed abnormal development suggesting that mortality occurs between seven and 35 days of development.

DISCUSSION

In a previous study, we proposed a candidate variant in the *SHBG* gene for embryonic lethality segregating with the MH1 haplotype in Montbeliarde cattle (Fritz *et al.*, 2013). Here, with the support of genotyping data we demonstrated that this variant is not recessive embryonic lethal. Subsequently, we fine mapped the embryonic lethal mutation in a 702 kb region on chromosome 19 and identified 27 variants in association with MH1. Among them, only the *PFAS* g.28511199C>T substitution was predicted to be deleterious. Moreover genotyping of more than 20,000 animals revealed a complete linkage between allele g.28511199T and the embryonic lethal mutation. In our initial study, we accidentally missed this strong candidate mutation since one of the two carrier bulls sequenced at that

time (ie. BOISLEVIN) had a no read coverage for this site. This compromised its genotype call for this SNP and misled our screen for potential causal mutations. These results demonstrate the importance of having substantial numbers of whole-genome sequenced individuals and of performing concordance tests to uncover causal variations when sequencing is performed at low to medium coverage. They also underline the high efficiency of large scale genotyping to statistically validate the causality of embryonic lethal variants.

PFAS g.28511199C>T variant is predicted to result in the substitution of an arginin which is entirely conserved among eukaryotes by a cystein in a Class I glutamine amidotransferase-like domain of the protein (PFAS p.R1205C). This perfect conservation reveals a strong evolutionary constraint and provides evidence that this residue plays an important role in the stability and/or activity of the protein**.** PFAS is involved in the fourth step of the *de novo* purine synthesis, a universal pathway amongst eukaryotes organisms (with the exception of some parasites) due to the central function of purine molecules in cells metabolisms such as DNA and RNA synthesis and energy molecules supply (Heinikoff *et al.*, 1987, Bønsdorff *et al.*, 2004). Although purine synthesis is a regulated balance between *de novo* and salvage pathways, it has been shown that *de novo* synthesis is more active in proliferating cells and therefore essential to meet high demand of purines for nucleic acids synthesis in early embryonic development (Alexiou *et al.*, 1992). Interestingly, mutations of several genes of this pathway have been reported to cause recessive embryonic lethality in different eukaryotic species (eg. *PRAT*, *GART*, *PAICS*; Clark, 1994; Ng *et al.*, 2009; Fritz *et al.*, 2013). Among them, we should highlight *GART*, in which we previously identified a missense variant (chr2:g.1277227A>C; p.N290T) associated with the embryonic lethal haplotype HH4 segregating in Holstein cattle (Fritz *et al.*, 2013). Concerning *PFAS*, several mutations have been described in mouse, causing dominant "short face" phenotype and most probably recessive embryonic lethality, since no homozygous mutant has ever been obtained from mating heterozygous animals (Palmer *et al.*, 2016). In addition, homozygous deficiency of *PFAS* ortholog gene *ade2* is lethal during pupal development in *Drosophila melanogaster* (Heinikoff *et al.*, 1986; Holland *et al.*, 2011). Taken together these arguments strongly support a causative role for the g.28511199C>T variant in MH1-associated embryonic lethality.

By analyzing the effect of this mutation on non-return rate at 56 days and by genotyping embryos form mating at risk, we demonstrated that mortality in homozygotes occurs between 7 and 35 days of gestation. At 7 days, homozygous mutant blastocysts were normal in appearance indicating that they had undergone a normal early embryonic development. Particularly, they were able to make the transition from maternal genome to zygotic genome expression, a critical phase happening around the 8 cells stage with a high nucleic acid demand for *de novo* mRNA transcription, and thus purines (Graf *et al.*, 2014). Interestingly, it has been previously demonstrated that maternal *PRAT*, *GART* and *PAICS* mRNA are important for early embryonic development and could attenuate the expression of severe phenotypes in homozygous mutant drosophila and zebra fish embryos (Malmanche and Clark, 2004, Ng

et al., 2009). Similarly, the presence of maternal *PFAS* mRNA and reserves in nucleic acid precursors could explain the survival of homozygous mutant embryos for a time after embryonic genome. It is also possible that the PFAS p.R1205C substitution leads to a decrease in protein activity and not to a total inactivation. In such situation, partial *PFAS* activity could be sufficient to meet purine requirements during early embryonic development but not during blastocyst elongation and trophoblast differentiation. Implantation of homozygous mutants and control embryos, close monitoring of gestations and gene expression analyses would be required to test these hypotheses and further understand molecular mechanisms causing the death of homozygous mutants.

Finally, mutations affecting genes belonging to the purine *de novo* and salvage synthesis pathways are investigated as potential candidates for severe metabolic disorders in humans, which are supposed to cause prenatal or early death in their most severe and hardly detectable forms (Baresova *et al.*, 2016). Interestingly, we have found in the human Exome Aggregation Consortium (ExAC) database the same p.R1205C substitution in the PFAS protein (GRCh37/hg19; Chr17:8172081C>T; Lek *et al.*, 2016), which segregates at a very low frequency in European and Asian populations (4 alleles over 117110 observed; no homozygote). Thus, the discovery of this recessive lethal candidate mutation in Montbeliard cattle provides an interesting model to investigate a potential effect of *PFAS* mutations in pregnancy and embryonic development in humans.

CONCLUSION

Combining genome scan for homozygous haplotype deficiency with whole genome sequence data analysis is a powerful approach for quickly identifying embryonic lethal mutations. However, it is of primary importance to statistically confirm the perfect linkage between the candidate mutations identified and the embryonic lethal phenotype by genotyping large populations. Here, after eliminating a previous candidate mutation in the *SHBG* gene, and analyzing new WGS data, we propose a strong candidate mutation (g.28511199C>T, PFAS p.R1205C) for recessive embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle. The absence of homozygous mutants among tens of thousands of genotyped animals, the perfect conservation of the affected residue among eukaryotes, and the critical function of PFAS and of the *de novo* purine synthesis pathway are convincing arguments of its embryonic lethal effect. This discovery provides the optimal tool to carry out an efficient selection against this embryonic lethal defect present at a high frequency in Montbeliarde and, to some extent, in Simmental dairy cattle.

ACKNOWLEDGMENTS

This study is part of the BOVANO project (ANR-14-CE19-0011) funded by the French Agence Nationale de la Recherche and APIS-GENE. The authors also acknowledge the financial contribution

of EU-FP7 FECUND grant (n°312097 - 2013-2017). P. Michot is recipient of a CIFRE PhD. Grant from ALLICE, with the financial support of ANRT and APIS-GENE. The authors are grateful to the partners of the 1000 bull genomes consortium for the excellent collaboration.

REFERENCES

Adams, H.A., T. Sonstegard, P.M. VanRaden, D.J. Null, C.P. Van Tassell, D.M. Larkin, and H.A. Lewin. 2016. Identification of a nonsense mutation in APAF1 that is likely causal for a decrease in reproductive efficiency in Holstein dairy cattle. J. Dairy Sci. 99: 6693-6701. http://dx.doi.org/10.3168/jds.2015-10517.

Alexiou, M. and H. J. Leese. 1992. Purine utilisation, *de novo* synthesis and degradation in mouse preimplantation embryos. Development. 114:185-192.

Baresova, V, M. Krijt, V. Skopova, O. Souckova, S. Kmoch, and M. Zikanova. 2016. CRISPR-Cas9 induced mutations along *de novo* purine synthesis in HeLa cells result in accumulation of individual enzyme substrates and affect purinosome formation. Mol. Genet. Metab. 119:270-277. http://dx.doi.org/10.1016/j.ymgme.2016.08.004.

Boichard, D., F. Guillaume , A. Baur, P. Croiseau, M.N. Rossignol, M.Y. Boscher, T. Druet , L. Genestout, J.J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. Anim. Prod. Sci. 52:115-120. http://dx.doi.org/10.1071/AN11119.

Bønsdorff, T., M. Gautier, W. Farstad, K. Rønningen, F. Lingaas, and I. Olsaker. 2004. Mapping of the bovine genes of the *de novo* AMP synthesis pathway. Anim. Genet. 35:438-444. http://dx.doi.org/10.1111/j.1365-2052.2004.01201.x.

Boussaha, M., D. Esquerré, J. Barbieri, A. Djari, A. Pinton, R. Letaief, G. Salin, F. Escudié, A. Roulet, S. Fritz, F. Samson, C. Grohs, M. Bernard, C. Klopp, D. Boichard, and D. Rocha. 2015. Genome-Wide Study of Structural Variants in Bovine Holstein, Montbéliarde and Normande Dairy Breeds. PLoS ONE 10:e0135931. http://dx.doi.org/10.1371/journal.pone0135931.

Boussaha, M., P. Michot, R. Letaief, C. Hoze, S. Fritz, C. Grohs, D. Esquerre, A. Duchesne, R. Philippe, V. Blanquet, F. Phocas, S. Floriot, D. Rocha, C. Klopp, A. Capitan, and D. Boichard. 2016. Construction of a large collection of small genome variations in French dairy and beef breeds using whole genome sequences. Genet. Sel. Evol. 48, 87. http://dx.doi.org/10.1186/s12711-016-0268-z.

Clark, D.V. 1994. Molecular and genetic analyses of Drosophila Prat, which encodes the first enzyme of *de novo* purine biosynthesis. Genetics. 136:547-57.

Daetwyler, H.D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R.F. Brondum, X. Liao, A. Djari, S.C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M.N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P.J. Bowman, D. Coote, A.J. Chamberlain, C. Anderson, C.P. VanTassell, I. Hulsegge, M.E. Goddard, B. Guldbrandtsen, M.S. Lund, R.F. Veerkamp, D.A.

Boichard, R. Fries, and B.J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46: 858-65. http://dx.doi.org/10.1038/ng.3034.

Fritz, S., A. Capitan, A. Djari, S.C. Rodriguez, A. Barbat, A. Baur, C. Grohs, B. Weiss, M. Boussaha, D. Esquerre, C. Klopp, D. Rocha, and D. Boichard. 2013. Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2. PLoS ONE 8: e65550. http://dx.doi.org/10.1371/journal.pone.0065550.

Graf, A., S. Krebs, V. Zakhartchenko, B. Schwalb, H. Blum, and E. Wolf. 2014. Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc. Natl. Acad. Sci. U. S. A. 111:4139-4144. http://dx.doi.org/10.1073/pnas.1321569111.

Henikoff, S. 1987. Multifunctional polypeptides for purine *de novo* synthesis. BioEssays 6 :8-13. http://dx.doi.org/10.1002/bies.950060104.

Henikoff, S., M.A. Keene, K. Fechtel, and J.W. Fristrom. 1986. Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands. Cell. 44: 33-42.

Holland, C., D.B. Lipsett and D.V. Clark. 2011. A Link Between Impaired Purine Nucleotide Synthesis and Apoptosis in Drosophila melanogaster. Genetics 188: 359-367. http://dx.doi.org/10.1534/genetics.110.124222.

Kumar, P., S. Henikoff, and P.C. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4:1073-1081. http://dx.doi.org/10.1038/nprot.2009.86.

Lek, M., K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A.H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, T. Tukiainen, D.P. Birnbaum, J.A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D.N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M.I. Kurki, A.L. Moonshine, P. Natarajan, L. Orozco, G.M. Peloso, R. Poplin, M.A. Rivas, V. Ruano-Rubio, S.A. Rose, D.M. Ruderfer, K. Shakir, P.D. Stenson, C. Stevens, B.P. Thomas, G. Tiao, M.T. Tusie-Luna, B. Weisburd, H.H. Won, D. Yu, D.M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J.C. Florez, S.B. Gabriel, G. Getz, S.J. Glatt, C.M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M.I. McCarthy, D. McGovern, R. McPherson, B.M. Neale, A. Palotie, S.M. Purcell, D. Saleheen, J.M. Scharf, P. Sklar, P.F. Sullivan, J. Tuomilehto, M.T. Tsuang, H.C. Watkins, J.G. Wilson, M.J. Daly, D.G. MacArthur, and E.A. Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285-291. http://dx.doi.org/10.1038/nature19057.

Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25: 1754-1760. http://dx.doi.org/10.1093/bioinformatics/btp324.

Malmanche, N. and D.V. Clark. 2004. *Drosophila melanogaster Prat*, a Purine *de novo* Synthesis Gene, Has a Pleiotropic Maternal-Effect Phenotype. Genetics 168: 2011–2023. http://dx.doi.org/10.1534/genetics.104.033134.

McClure, M.C., D. Bickhart, D. Null, P. VanRaden, L.Y. Xu, G. Wiggans, G. Liu, S. Schroeder, J. Glasscock, J. Armstrong, J.B. Cole, C.P. Van Tassell, and T.S. Sonstegard. 2014. Bovine Exome Sequence Analysis and Targeted SNP Genotyping of Recessive Fertility Defects BH1, HH2, and HH3 Reveal a Putative Causative Mutation in SMC2 for HH3. Plos ONE, 9: e92769. http://dx.doi.org/10.1371/journal.pone.0092769.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing nextgeneration DNA sequencing data. Genome Res. 20: 1297–1303. http://dx.doi.org/10.1101/gr.107524.110.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069-2070. http://dx.doi.org/ 10.1093/bioinformatics/btq330.

Menzi, F., N. Besuchet-Schmutz, M. Fragniere, S. Hofstetter, V. Jagannathan, T. Mock, A. Raemy, E. Studer, K. Mehinagic, N. Regenscheit, M. Meylan, F. Schmitz-Hsu, and C. Drogemuller C. 2016. A transposable element insertion in APOB causes cholesterol deficiency in Holstein cattle. Anim. Genet. 47: 253-257. http://dx.doi.org/10.1111/age.12410.

Muñoz, M, A. Uyar, E. Correia, C. Ponsart, C. Guyader-Joly, D. Martínez-Bello D, B. Marquant-Le Guienne, A. Fernandez-Gonzalez, C. Díez, J. N. Caamaño, B. Trigal, P. Humblot, S. Carrocera, D. Martin, E. Seli, and E. Gomez. 2014. Metabolomic prediction of pregnancy viability in superovulated cattle embryos and recipients with fourier transform infrared spectroscopy. Biomed. Res. Int. 2014:e608579. http://dx.doi.org/10.1155/2014/608579.

Ng, A., R.A. Uribe, L. Yieh, R. Nuckels, and J.M. Gross. 2009. Zebrafish mutations in gart and paics identify crucial roles for *de novo* purine synthesis in vertebrate pigmentation and ocular development. Development 136:2601-2611. http://dx.doi.org/10.1242/dev.038315.

Palmer, K., H. Fairfield, S. Borgeia, M. Curtain, M.G. Hassan, L. Dionne, S. Yong Karst, H. Coombs, R.T. Bronson, L.G. Reinholdt, D.E. Bergstrom, L.R. Donahue, T.C. Cox, and S.A. Murray. 2016. Discovery and characterization of spontaneous mouse models of craniofacial dysmorphology. Dev. Biol. 415: 216-227. http://dx.doi.org/10.1016/j.ydbio.2015.07.023.

Pausch, H., H. Schwarzenbacher, J. Burgstaller, K. Flisikowski, C. Wurmser, S. Jansen, S. Jung, A. Schnieke, T. Wittek, and R. Fries. 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. BMC Genomics. 16: 312. http://dx.doi.org/10.1186/s12864-015-1483-7.

Reinartz, S. and O. Distl. 2016. Validation of Deleterious Mutations in Vorderwald Cattle. PLoS ONE 11: e0160013. http://dx.doi.org/10.1371/journal.pone.0160013.

Robinson, J.T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, and J.P. Mesirov. 2011. Integrative genomics viewer. Nature Biotechnology. 29: 24–26. http://dx.doi.org/10.1038/nbt.1754.

Rozen S. and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-86.

Sahana, G., U.S. Nielsen, G.P. Aamand, M.S. Lund, and B. Guldbrandtsen. 2013. Novel Harmful Recessive Haplotypes Identified for Fertility Traits in Nordic Holstein Cattle. Plos ONE 8: e82909. http://dx.doi.org/10.1371/journal.pone.0082909.

Sargolzaei, M., J.P. Chesnais, and F.S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 15: 478. http://dx.doi.org/10.1186/1471-2164-15-478.

Schutz, E., C. Wehrhahn, M. Wanjek, R. Bortfeld, W.E. Wemheuer, J. Beck, and B. Brenig. 2016. The Holstein Friesian Lethal Haplotype 5 (HH5) Results from a Complete Deletion of TBF1M and Cholesterol Deficiency (CDH) from an ERV-(LTR) Insertion into the Coding Region of APOB. Plos ONE 11: e0154602. http://dx.doi.org/10.1371/journal.pone.0154602.

Schwarzenbacher, H., J. Burgstaller, F.R. Seefried, C. Wurmser, M. Hilbe, S. Jung, C. Fuerst, N. Dinhopl, H. Weissenbock, B. Fuerst-Waltl, M. Dolezal, R. Winkler, O. Grueter, U. Bleul, T. Wittek, R. Fries, and H. Pausch. 2016. A missense mutation in TUBD1 is associated with high juvenile mortality in Braunvieh and Fleckvieh cattle. BMC Genomics 17: 400. http://dx.doi.org/10.1186/s12864-016-2742-y.

Sonstegard, T.S., J.B. Cole, P.M. VanRaden, C.P. Van Tassell, D.J. Null, S.G. Schroeder, D. Bickhart, and M.C. McClure. 2013. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. PLoS ONE 8: e54872. http://dx.doi.org/10.1371/journal.pone.0054872.

VanRaden, P.M., K.M. Olson, D.J. Null, and J.L. Hutchison. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J. Dairy Sci. 94:6153–6161. http://dx.doi.org/10.3168/jds.2011-4624.

Venhoranta, H., H. Pausch, K. Flisikowski, C. Wurmser, J. Taponen, H. Rautala, A. Kind, A. Schnieke, R. Fries, H. Lohi, and M. Andersson. 2014. In frame exon skipping in UBE3B is associated with developmental disorders and increased mortality in cattle. BMC Genomics. 15:890. http://dx.doi.org/10.1186/1471-2164-15-890.

Weckx, S., J. Del-Favero, R. Rademakers, L. Claes, M. Cruts, P. De Jonghe, C. Van Broeckhoven, and P. De Rijk. 2005. novoSNP, a novel computational tool for sequence variation discovery. Genome Res. 15: 436-42. http://dx.doi.org/10.1101/gr.2754005.

Ye, K., H. M. Schulz, Q. Long, R. Apweiler, and Z. Ning. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865-71. http:// dx.doi.org/ 10.1093/bioinformatics/btp394.

Table 1. Association between MH1 status and SHBG p.Q52X genotype

	+/+	MH1/+	MH1/MH1	Total
SHBG Q52/Q52	38233	3	0	38236
SHBG Q52/X52	3256	6137	0	9393
SHBG X52/X52	13	229	0	242
Total	41502	6369	0	47871

+/+: Mh1 Non-carriers; +/MH1 and MH1/MH1: respectively MH1 heterozygous and homozygous carriers.

Table 2. Genotypes of 20,350 Montbeliard animals for the PFAS variant on the EuroG10K SNP BeadChip

Genotypes	Number Observed	Number Expected
PFAS R1205/R1205	17427	17532
PFAS R1205/C1205	2923	2713
PFAS C1205/C1205	0	105

Table 3. Loss in non-return rate in mating at risk

Mating at risk	Category	Number of matings at risk	Proportion of mating at risk compared to all matings (%)	Loss in NRR56
Carrier x Daughter of a carrier	Heifer	76 723	4,06	-4,77
	Cow	193 390	4,51	-4,94
Carrier x Carrier	Heifer	774	1,49	-11,15
	Cow	797	1,38	-12,58

All losses in NRR56 are significant with P<0.05

Table 4. Genotypes of seven days old embryos produced from carrier parents

Genotype	Embryos observed	Embryos expected supposing recessive embryonic lethality prior 7 days	Embryos expected supposing no recessive embryonic lethality prior 7 days
CC	2	6	4
CT	11	11	9
TT	4	0	4
Total	17	17	17

Michot *et al.*, Figure 1

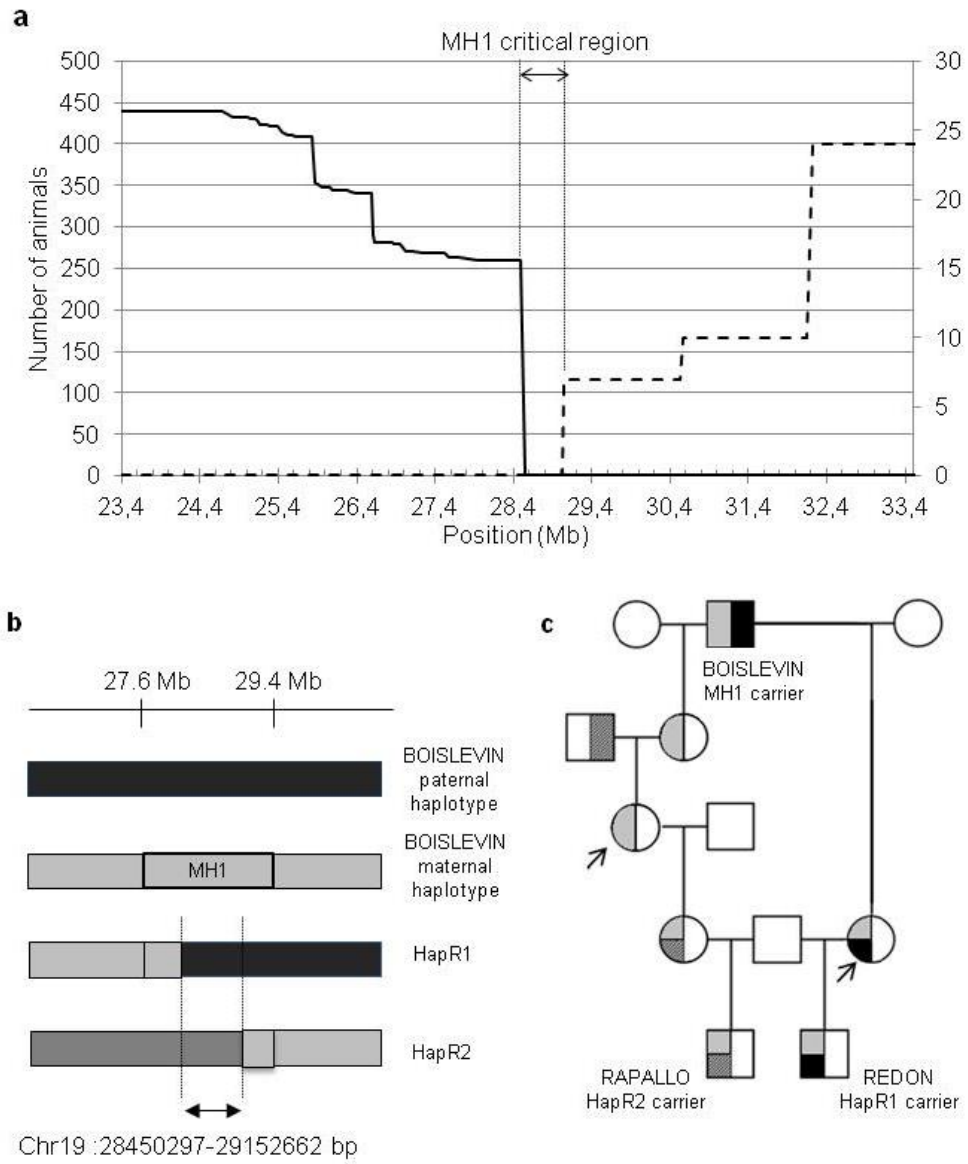

Figure 1. Fine mapping of the MH1-associated embryonic lethal mutation

a. Identification of recombinant haplotypes in the descendants of BOISLEVIN using Illumina BovineSNP50 genotyping data. The black line (respectively dotted line) corresponds to the number of descendants of BOISLEVIN that are homozygous for large identical by descent segments on the left (respectively right) side of MH1. Most of these animals carry one copy of the original segment containing MH1 and a recombining haplotype. **b.** Diagram of recombining haplotypes HapR1 and HapR2. These were used to define the critical interval in which the MH1-associated embryonic lethal mutation should be located. **c.** Pedigree of the two bulls that have spread HapR1 and HapR2 recombining haplotypes in the Montbéliarde population. Arrows indicate the first supposed carriers of these recombining haplotypes.

Michot *et al.*, Figure 2

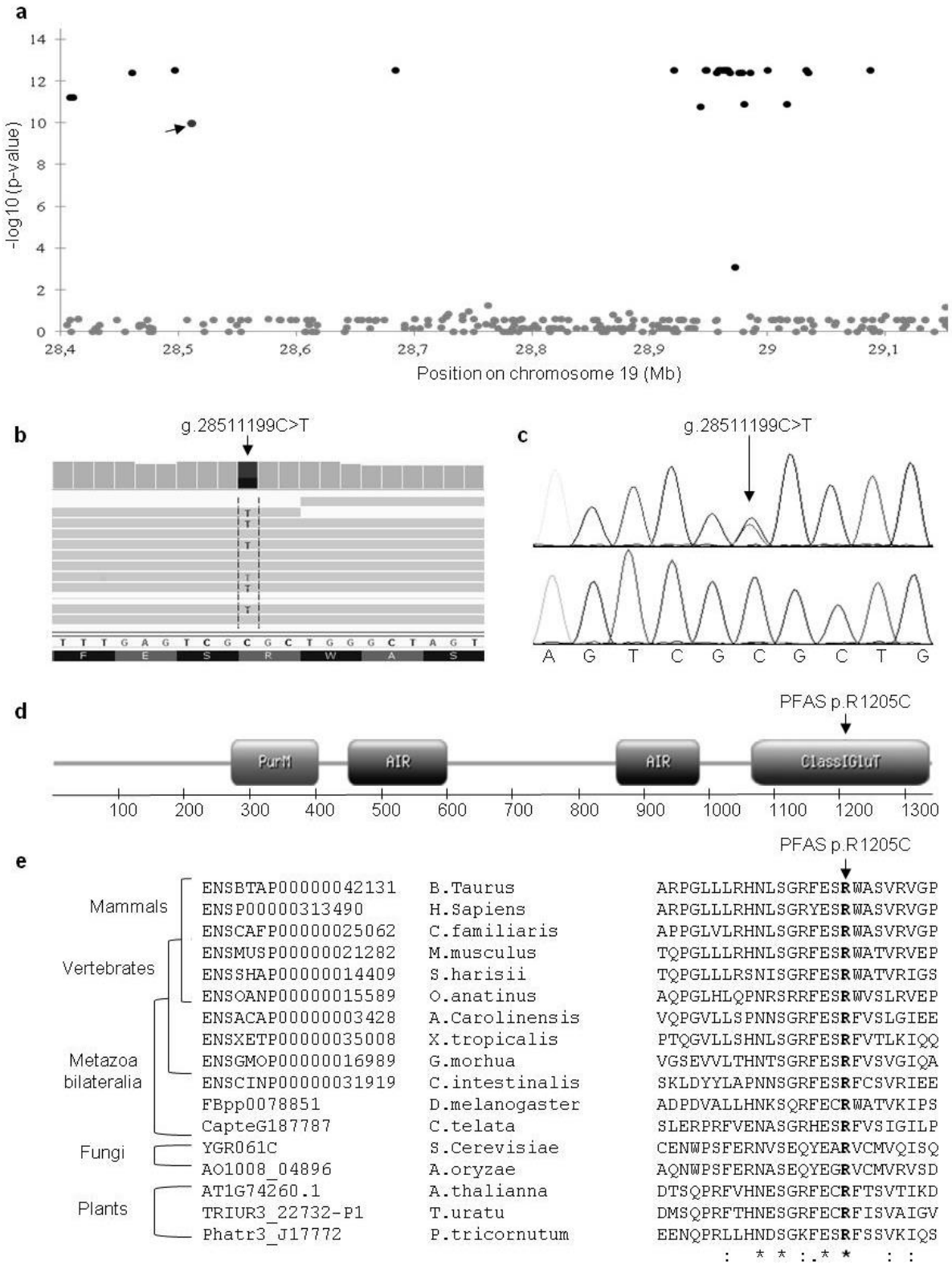

Figure 2. Identification of a missense variant in PFAS associated with the MH1 embryonic lethal haplotype

a. Association of filtered polymorphisms with MH1 haplotype in the 702 kb critical region. Variants with a significant p-value (<0.05) are represented in black. The arrow indicates the *PFAS* deleterious candidate variant (g.chr19:28511199C>T). Details on variants functional annotations are provided in Supplementary table 1. **b**. IGV (integrative genome viewer) screenshot showing sequenced reads supporting the g.28511199C>T substitution on chromosome 19 in one heterozygous carrier bull. **c**. Electrophoregrams from a heterozygous carrier (up) of the g.28511199C>T substitution on chromosome 19 and from a wild type animal (down). This substitution, located in exon 26 of *PFAS*, is predicted to result in a p.R1205C amino-acid substitution at the protein level. **d**. Diagram of the PFAS protein domains: a PurM-like N-terminal domain (PurM), two AIR synthase-related protein C-terminal domains (AIR) and the Class I glutamine amidotransferase-like domain (ClassIGluT). The p.R1205C amino-acid change is located in a ClassIGluT domain. PFAS protein features and domains positions were retrieved from Ensembl and Uniprot. **e.** Alignment of PFAS orthologs in different species showing a complete conservation of the p.R1205 amino-acid among eukaryotes.

APPENDIX

Supplementary table 1. Statistics and annotations regarding sequence variants that are the most associated with the MH1-associated embryonic lethal mutation.

ID: entry code of the sequence variants in dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/); Conservation: yes indicates that the reference allele is located within a constrained element among 40 eutherian mammals genomes according to Ensembl (http://www.ensembl.org/; EPO_LOW_COVERAGE) and that this nucleotide is well conserved among species; *: the two different alleles C and T are observed in various species (e.g. bos taurus, ovis aries and oryctolagus cuniculus for allele C, and Homo sapiens, Mus musculus, and Canis lupus for allele T).

CHAPITRE 5 : APPROCHES DE GENETIQUE INVERSE APPLIQUEES A LA DETECTION DE VARIANTS DELETERES RARES.

La mise en place d'observatoires nationaux pour détecter les émergences d'anomalies génétiques et l'utilisation des technologies de génotypage et de séquençage à haut débit se sont révélées très efficaces pour cartographier et identifier en un temps record nombre de mutations causales. Toutefois cette approche classique, du phénotype au gène, repose sur la description clinique la plus précise possible d'animaux jugés anormaux et la disponibilité d'échantillons biologiques. Lorsque les phénotypes ne sont pas facilement disponibles, cette approche ne peut être mise en œuvre. C'est le cas par exemple quand les phénotypes sont très peu spécifiques et souvent confondus avec d'autres pathologies courantes en élevage (p. ex. diarrhée, déficit immunitaire,…) ou bien quand ils sont peu visibles dans les conditions d'élevage courantes (surdité, cécité,…). C'est également le cas pour les anomalies responsables de morts précoces au cours de la gestation, qui se traduisent indirectement par une baisse de fertilité à l'échelle de la population (voir *Chapitre 4*). Ces anomalies qui passent inaperçues sont sans doute assez fréquentes même s'il est difficile d'en apprécier le nombre, du fait du manque d'information.

Pour rechercher et caractériser ces anomalies peu visibles, la stratégie doit donc être différente. Au lieu de partir des cas pour remonter au gène, on peut imaginer d'analyser la séquence du génome d'individus afin de détecter des mutations candidates sur la base de leur annotation et en confirmer le caractère délétère sur des animaux ciblés, selon une approche dite de « génétique inverse ». Dans ce chapitre, nous présentons comment une telle approche a été mise en œuvre dans le cadre de ma thèse.

La stratégie d'approche inverse que nous avons mise en place comprend trois étapes (Figure 8):

i) l'identification de variants avec un fort potentiel délétère à partir des données de séquence,
ii) le génotypage de ces candidats à grande échelle dans nos populations bovines afin de sélectionner les plus intéressants et de repérer des animaux d'intérêt et enfin,
iii) la caractérisation de l'effet délétère de ces mutations à travers le suivi d'accouplements entre porteurs et/ou le phénotypage fin des animaux homozygotes génotypés.

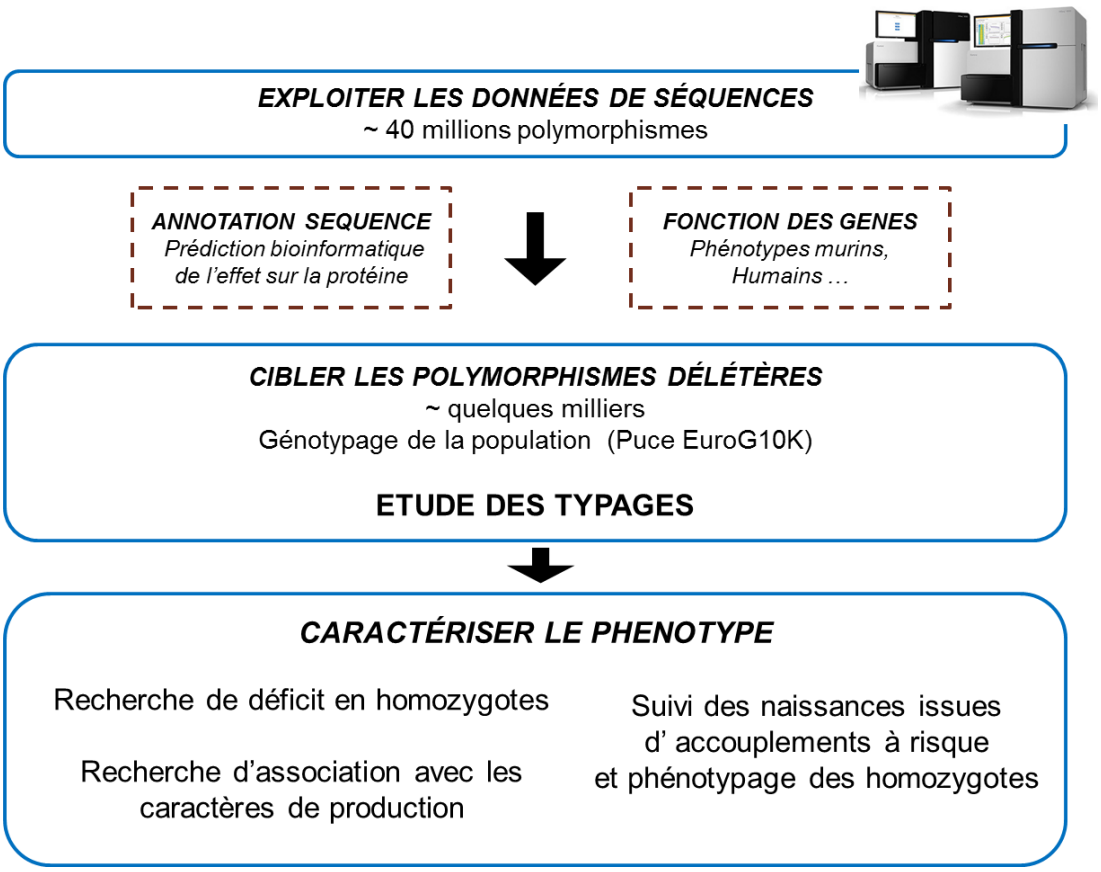

Figure 8 : Stratégie de génétique inverse appliquée à la détection de mutations délétères chez le bovin

Nous présenterons ici les résultats et conclusions préliminaires de l'étude. Suite au retard de mise en production de la version 4 (v4) de la puce EuroG10K, l'étape de caractérisation fonctionnelle des variants candidats n'a pu être réalisée que partiellement.

I. Matériels et méthodes

A. Exploitation des données de séquence de génomes complets

1. Données de séquences, détection et annotation des variants

Pour cette étude, nous avons exploité les séquences de génomes complets de 429 taureaux d'IA de 15 races différentes mises à disposition par le consortium 1000 génomes bovins (« 1000 bull genomes project ») ainsi que de 115 taureaux d'IA des trois principales races laitières françaises (Holstein, Normande et Montbéliarde) séquencés dans le cadre des projets REPROSEQ et CARTOSEQ. CARTOSEQ est un projet financé par l'ANR et APIS-GENE, coordonné par l'équipe G2B, et dont

l'objectif était la caractérisation de QTL détectés par analyse d'association à grande échelle. Dans ce projet, les génomes de 67 animaux ont été séquencés. REPROSEQ est un projet financé par APIS-GENE et coordonné par Allice dont l'objectif visait à mettre en évidence l'architecture génétique de caractères de reproduction par une approche intégrée de cartographie et d'études d'expression. Ce projet a conduit au séquençage du génome de 48 animaux.

Ces animaux font partie des principaux contributeurs de leurs races et, de ce fait, sont porteurs d'une grande partie des variations génétiques qui ségrégent aujourd'hui dans ces populations bovines.

Les séquences ont été obtenues à partir de librairies pair-end séquencées sur plateformes Illumina Hiseq 2000 et 2500. Les lectures générées, de 100 paires de bases en moyenne, ont par la suite été alignées à l'aide du logiciel BWA (Li et Durbin, 2009) sur l'assemblage de référence UMD3.1 du génome bovin. La couverture moyenne obtenue est d'environ 10x par animal. La détection des variants (SNP et petits INDELs) a été réalisée de façon indépendante dans chaque base de données à l'aide des logiciels *Samtools* (Li *et al.*, 2009) pour les séquences du consortium 1000 génomes bovins et *Genome Analysis Tool Kit* (GATK) (McKenna *et al.*, 2010) pour les données de séquence françaises. Les variants ont été annotés avec le logiciel *Variant Effet Predictor* (Ve!P) d'Ensembl (McLaren *et al.*, 2016) et les effets des substitutions d'acides aminés ont été estimés par le logiciel SIFT (Kumar *et al.*, 2009).

2. Sélection de variants délétères race-spécifiques

Nous avons cherché à identifier des variants avec un effet potentiellement délétère et ségrégant dans des races pour lesquelles des génotypes sont réalisés en routine en France dans le cadre de la sélection génomique. Ceci a restreint l'étude aux grandes races laitières françaises Montbéliarde, Normande et Holstein ainsi que Brune, dont la sélection génomique a été mise en place récemment, et à deux grandes races allaitantes françaises, la Charolaise et la Limousine. Nous nous sommes également limités la détection de variants sur les 29 autosomes afin de restreindre le champ de recherche à la détection d'anomalies autosomales récessives.

Les données de séquence disponibles ont été filtrées pour sélectionner des SNP et petits InDels bi-alléliques pour lesquels aucun individu homozygote pour l'allèle alternatif n'a été observé et qui ségrégent dans une seule race du jeu de données étudié. Dans le cas de la race Montbéliarde, les animaux Simmental (N=87) ont été écartés lors de l'application du filtre race-spécifique puisque ces deux races n'ont divergé que récemment. Ainsi, les variants ségrégant à la fois en Montbéliarde et en Simmental n'ont pas été exclus de l'étude. Ensuite nous avons retenu les variants prédits comme ayant un effet délétère sur les protéines, c'est-à-dire les : (i) gains ou pertes d'un codon stop ; (ii) modifications du codon initiateur de traduction ; (iii) modifications des sites accepteurs ou donneurs d'épissage ; (iv) insertions ou délétions induisant un décalage du cadre de lecture lors de la traduction ; (v) substitutions d'acides aminés prédites comme délétères par le logiciel SIFT (score <0.05) ; et enfin (vi) les insertions et délétions d'acides aminés conservant le cadre de lecture (donc multiples de 3 bases).

Pour limiter les artefacts de séquençage, nous avons exclu les variants avec un score QUAL (qualité de détection du variant) inférieur à 30.

3. Annotations complémentaires

En plus des annotations Ve!P nous avons eu recours à plusieurs bases de données pour prédire l'effet biologique de ces variants. Ainsi, nous nous sommes appuyés sur la base d'annotation de phénotypes murins *Mammalian Phenotype Ontologies* accessible en ligne à partie de la base de données du *Jackson Laboratory* (*Mouse Genome Database,* www.informatics.jax.org), la base des syndromes génétiques humains *Online Mendelian Inheritance in Man* (OMIM) développée par l'université John Hopkins (www.omim.org) ; ainsi que la base des syndromes génétiques chez les animaux domestiques *Online Mendelian Inheritance in Animals* (OMIA) de l'université de Sydney (http://omia.angis.org.au).

4. Choix des variants

Le premier filtre appliqué aux données de séquence a isolé plusieurs milliers de variants. Tous ne pouvant être intégrés à la puce de génotypage Bovine SNP EuroG10K d'Illumina et testés, nous avons opéré un choix arbitraire consistant à sélectionner en priorité les types de variants les plus délétères (gain de codon stop, décalage du cadre de lecture, sites d'épissage), ainsi que les phénotypes les plus facilement caractérisables chez les animaux homozygotes et impactant le plus la filière. Nous avons ainsi porté plus particulièrement notre attention sur des gènes dont la mutation est responsable chez l'homme ou les espèces modèles d'anomalies de la reproduction ; de mortalités embryonnaires, peri ou post-partum ; ou encore de symptômes dégénératifs d'apparition plus ou moins tardive. Malgré tout, nous avons également conservé quelques variants affectant des gènes de fonction mal connue mais appartenant à des familles ou voies de gènes possédant des fonctions biologiques importantes pour le développement, le métabolisme ou le maintien de l'organisme afin de participer à l'annotation fonctionnelle du génome bovin.

B. Génotypage à grande échelle sur la population

1. Préparation des séquences de variants pour intégration sur la puce de génotypage Illumina EuroG10K

Les séquences d'ADN de 100 pb de chaque côté du variant ont été récupérées à partir de l'assemblage de référence bovin UMD3.1 sur le *genome browser UCSC* (http://genome-euro.ucsc.edu), pour dessiner les sondes d'ADN permettant le typage des SNP sur la puce Illumina. Chaque sonde est composée d'une séquence de 50 bases spécifiques du variant, flanquant et se terminant par le variant biallélique étudié. De cette façon, les échantillons d'ADN des animaux à génotyper s'hybrident avec l'une, l'autre ou les deux versions de la sonde en fonction des allèles portés. Ce procédé est particulièrement adapté aux

variants de type SNP. Il peut aussi être adapté aux insertions et délétions en testant la première base différant entre allèles au point de cassure, mimant ainsi un variant SNP, en imposant le côté commun pour l'amorce. Cette stratégie est possible dès lors qu'une séquence unique et non polymorphe jouxte le variant testé.

2. Calibration des clusters de génotypes

Les génotypes des individus ont été déterminés à l'aide du logiciel d'analyse *Genome Studio®*. Il permet de déterminer automatiquement un génotype pour chaque individu à partir des valeurs d'intensité de luminescence obtenues en fonction de l'hybridation spécifique de l'ADN aux sondes qui définissent les deux allèles du SNP. Pour un SNP donné, le logiciel réalise une classification (ou clusters) des échantillons avec des valeurs d'intensités similaires aux mêmes longueurs d'onde et leur attribue un génotype identique.

Selon les recommandations d'Illumina, une étape de calibration est systématiquement réalisée à l'arrivée de chaque nouvelle version d'une puce de génotypage en laboratoire. Elle permet de vérifier la qualité de génotypage de chaque SNP et de définir la position des trois clusters possibles appliquée automatiquement pour la détermination des génotypes sur les analyses en routine.

Cette étape a été réalisée à partir d'un panel d'individus (N=384) représentatifs des trois génotypes pour la plupart des marqueurs usuels de la puce. L'algorithme du logiciel *Genome Studio®* est appliqué dans un premier temps pour définir une première position automatique. Différents points de contrôle sont évalués, et en particulier le *call rate* par SNP qui donne le nombre de résultats de génotypes obtenus sur le nombre d'animaux analysés. Si cette valeur est trop faible, elle traduit un défaut de détermination des génotypes des individus à partir des valeurs d'intensité enregistrées. La cause peut être une mauvaise hybridation de la sonde avec l'ADN de l'individu ou bien une mauvaise estimation de la position des clusters. Lors de la calibration, les SNP avec un *call rate* trop faible sont retravaillés individuellement pour définir de meilleurs clusters ou bien les écarter de la suite des analyses si aucune amélioration ne peut être apportée. En général, les SNP pour lesquels le *call rate* est inférieur à 95% après cette étape ne sont pas conservés.

Contrairement aux marqueurs standards de la puce EuroG10K, les variants sélectionnés dans cette étude sont rares et peu représentés dans le panel d'individus tests. De ce fait, les clusters définis automatiquement pour ces variants ne sont pas toujours fiables. Ceci entraîne un fort taux d'élimination de SNP d'intérêt pour différents projets de recherche à la suite des contrôles qualité appliqués par le laboratoire de génotypage. Pour résoudre ce problème, une seconde étape de calibration des clusters a été appliquée à partir d'un set beaucoup plus large de 26 400 individus appartenant à 15 races différentes. Seuls les SNP pour lesquels aucun résultat de génotype n'a été rendu ont été retravaillés manuellement, soit dans cette étude 364 variants. Nous avons porté une attention particulière aux SNP pour lesquels l'un des génotypes était absent. En effet, l'algorithme recherche par défaut trois génotypes et est capable

d'estimer la position d'un cluster dans le cas où le génotype serait manquant. Cependant, ces SNP sont souvent associés à des erreurs d'attribution des génotypes par décalage du cluster des animaux hétérozygotes. Inversement, si le groupe des homozygotes alternatifs existe mais est décalé par rapport à sa position théoriquement attendue, le logiciel ne les repère pas et rend un résultat nul ou hétérozygote pour ces individus.

Les SNP pour lesquels aucune amélioration ne pouvait être apportée ont été identifiés et écartés de la suite des analyses. Une dernière vérification a été appliquée sur le *call rate* par SNP recalculé sur le jeu d'individus utilisé pour travailler les clusters. De nouveau, les variants associés à chaque race avec un *call rate* toujours inférieur à 95% ont été écartés.

3. Génotypes utilisés

L'EuroG10K est la puce utilisée en routine pour la sélection génomique en France (Boichard *et al.*, 2012b). Les génotypes ont été récupérés pour tous les animaux génotypés dans ce cadre. Il s'agit principalement de jeunes animaux de moins d'un an.

Pour des raisons techniques, les clusters de génotypes définis sur la version 4 de l'EuroG10K ne sont pas applicables pour le nouveau design de cette puce (version 5) utilisée depuis janvier 2016. C'est pourquoi nous avons pris en compte uniquement les résultats de typages réalisés sur la version 4. Les races Normande, Montbéliarde et Holstein comptent respectivement 8 257, 23 248 et 34 097 analyses réalisées. Les races Limousine, Charolaise et Brune ont des effectifs plus réduits avec respectivement 247, 975 et 256 typages réalisées. Ces effectifs correspondent au nombre total de typages réalisés sur la période janvier-octobre 2015 au laboratoire Labogena, le seul laboratoire réalisant le décodage des marqueurs recherche de l'INRA.

4. Calculs de fréquences et statistiques appliquées aux résultats de génotypage

Les fréquences alléliques et génotypiques ont été calculées pour chaque SNP et dans chaque race étudié(e). Le nombre attendu d'animaux à chaque génotype a été estimé selon la loi de Hardy-Weinberg (HW) ; soit pour deux allèles A, B aux fréquences respectives p et q, et pour N individus génotypés les fréquences des homozygotes AA et BB sont de p^2N et q^2N respectivement, et la fréquence des hétérozygotes est de 2pqN. En première approche, l'effectif observé a été comparé à l'effectif théorique par un test de χ^2, avec un seuil de significativité α fixé à 5%.

II. Résultats et discussions

A. Variants délétères races-spécifiques après filtre des données de séquences

A partir des données de séquence, nous avons isolé 9374 variants potentiellement délétères, spécifiques à l'une des six races étudiées et pour lesquels aucun homozygote à l'allèle alternatif n'a été identifié parmi l'ensemble des animaux séquencés (Tableau 5). La majorité de ce set est constitué de SNP (entre 92 et 100% des variants choisis en fonction de la race). Le nombre de variants par race augmente en fonction du nombre d'individus séquencés. En effet, plus on dispose d'individus séquencés, plus on capte les variations rares du génome. Ce nombre varie de 327 variants pour la race Charolaise avec le plus petit effectif d'animaux séquencés (N=8) à près de 6000 variants pour la race Holstein avec l'effectif le plus élevé (N=120). Plus de 65% (N=6153) des variants isolés ont été identifiés chez un seul porteur et jusqu'à 38 individus pour les variants restants (Figure 9). Sur les 491 animaux séquencés, les fréquences alléliques sont inférieures à 5% et même inférieures à 1% pour 98% des variants sélectionnés. Cette base de variants est donc constituée en majorité de variants rares chez le bovin.

En ce qui concerne leur nature, ces variants délétères se répartissent comme suit : 82% (N=7685) de substitutions d'acides aminés, 9,3% (N=871) de codons stop, 5,5% (N=511) de sites donneurs ou accepteurs d'épissage et 2,2% (N=203) d'insertions ou délétions induisant un décalage du cadre de lecture (Tableau 6). Le reste est constitué de variants affectant le codon initiateur ou le codon stop de la protéine et d'insertions ou délétions d'acides aminés sans décalage du cadre de lecture de la synthèse protéique.

Tableau 5 : Variants races-spécifiques sans homozygotes pour l'allèle alternatif dans les données de séquence

	Brune	Charolaise	Holstein	Limousine	Montbéliarde	Normande	**Total**
Nombre d'individus séquencés	43	8	120	25	23	19	**238**
Nombre d'INDEL (proportion)	75 (7,3%)	1 -	133 (2,3%)	10 (1,1%)	59 (7,9%)	32 (7,5%)	**310 (3.3%)**
Nombre de SNP (proportion)	942 (92,7%)	326 (100%)	5831 (97,7%)	885 (98,9%)	683 (92 ,1%)	397 (92.5%)	**9064 (96.7%)**
Total	**1017**	**327**	**5964**	**895**	**742**	**429**	**9374**

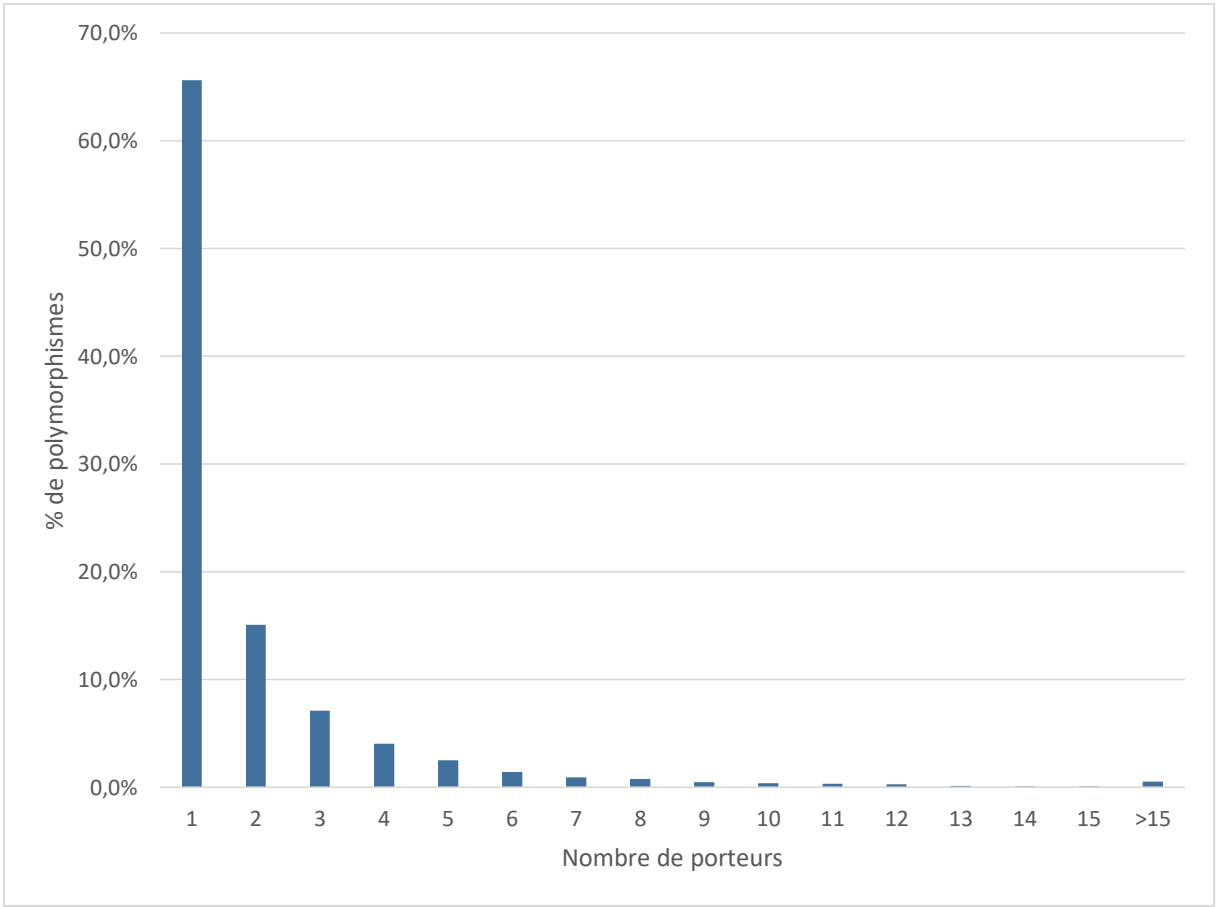

Figure 9: Distribution des variants en fonction du nombre d'animaux porteurs

Tableau 6. Distribution des variants hétérozygotes selon leur conséquence sur la protéine.

	Brune	Charolaise	Holstein	Limousine	Montbéliarde	Normande	Total
Substitution délétère	871	287	4803	758	599	367	7685
Gain d'un codon stop	49	25	666	59	54	18	871
Modification d'un site d'épissage Accepteur	9	5	135	31	19	7	206
Donneur	21	7	227	36	10	4	305
InDel modifiant le cadre de lecture	43	1	85	9	45	20	203
InDel conservant le cadre de lecture	19		30	1	10	10	70
Modification du codon initiateur	3	2	10		3	2	20
Perte du codon stop	2		8	1	2	1	14
Total	**1017**	**327**	**5964**	**895**	**742**	**429**	**9374**

Au total, 6433 gènes sont affectés par au moins un variant, certains pouvant être concernés par plusieurs variants identifiés dans une ou plusieurs races. Environ 14,5% (N=935) de ces gènes codent pour une protéine non caractérisée ou un récepteur olfactif, soit 1306 variants sur les 9374 filtrés (entre 10 et 30% des variants hétérozygotes de chaque race). Le génome bovin contient environ un millier de séquences identifiées correspondant à des récepteurs olfactifs (Lee *et al.*, 2013) qui contiennent souvent de nombreuses variations avec un enrichissement en mutations délétères (Das *et al.*, 2015). Le nombre de variations dans ces familles de récepteurs olfactifs que nous identifions en ciblant les variants délétères semble cohérent.

Les gènes codant pour des protéines non caractérisées ou des récepteurs olfactifs possèdent peu d'informations phénotypiques et fonctionnelles qui permettraient une comparaison avec d'autres espèces. De façon similaire à ce qui se pratique chez la souris par exemple, l'étude de ces variations par une approche inverse pourrait permettre d'améliorer l'annotation de la fonction de ces gènes. Cependant cela suppose un dispositif de phénotypage complet et très large sur les animaux. Nous avons fait le choix d'écarter les variants concernés pour la suite de l'analyse afin de faciliter la recherche de phénotypes.

Dans la liste des gènes affectés par nos variants et dont la fonction est connue, 1045 gènes sont associés à au moins un syndrome génétique humain (*www.omim.org*) et 159 gènes sont associés à au moins une anomalie génétique chez les espèces domestiques (*www.omia.org*). Cinquante-deux gènes (98 variants) ont déjà été associés à un syndrome génétique chez le bovin. Cependant, seulement quelques variants correspondent aux mutations candidates identifiées dans ces gènes.

Ainsi, nous avons retrouvé les mutations associées aux QTL de mortalité embryonnaire HH1, HH3 et HH4 en race Holstein (respectivement *APAF1* p.Q581X, *SMC2* p.F1305S et *GART* p.N290T), ainsi que

les mutations *SHBG* p.Q52X et *PFAS* p.R1205C associées au QTL de mortalité embryonnaire MH1 en race Montbéliarde (Fritz *et al.*, 2013 ; Daetwyler *et al.*, 2014 ; Michot *et al.*, submitted). Nous avons également retrouvé les mutations causales impliquées dans deux syndromes récessifs létaux : *OPA3* p.Q115X responsable d'une cardiomyopathie en race Holstein (Owczarek-Lipska *et al.*, 2011) et la substitution *SPAST* p.R560Q impliquée dans le syndrome de démyélinisation de la moelle épinière (SMD) en race Brune (Thomsen *et al.*, 2010) et enfin en race Charolaise, la mutation d'intérêt agronomique dans le gène de la myostatin (*MSTN* p.Q204X) associée à l'hypertrophie musculaire (Grobet *et al.*, 1998).

A l'exception de l'hypertrophie musculaire, toutes ces mutations sont récessives, race-spécifiques et génèrent un phénotype létal. Les retrouver « à l'aveugle » parmi les taureaux fondateurs séquencés, dont on connait le statut porteur, constitue un point de contrôle positif pour le filtre que nous avons appliqué aux données.

Cependant, nous nous attendions à détecter également les mutations pour au moins deux autres anomalies récessives race-spécifiques : les mutations dans les transporteurs membranaires *SCL35A3* (g.chr3:43,412,427C>A ; p.V180F) et *SLC37A2 (*g.chr29:28,879,810C>T ; p.R12X) responsables respectivement des syndromes de *complex vertebral malformation* (CVM) en race Holstein (Thomsen *et al.*, 2006) et de mortalité embryonnaire associée à l'haplotype MH2 en race Montbéliarde (Fritz *et al.*, 2013). Nous sommes retournés aux données de séquences et avons identifié deux taureaux porteurs de ces mutations dans d'autres races : un taureau Normand porteur de la mutation associée à MH2 et un taureau Angus porteur de la mutation associée à CVM. Pour le taureau de race Normande, dont nous disposions des fichiers bam, nous avons vérifié la qualité de couverture de cette position à l'aide du logiciel de visualisation graphique *Integrative Genome Viewer* (Figure 10). La visualisation de la mutation sous IGV montre qu'une seule lecture contre dix supporte l'allèle de référence et que le variant est identifié sur la fin de cette lecture. Ces observations confirment clairement que le variant identifié par GATK pour ce taureau Normand est un artéfact de séquençage. La même vérification n'a pas pu être effectuée pour CVM. Le taureau de race Angus peut également avoir été déclaré porteur du fait d'une erreur de séquençage. Il faut aussi envisager qu'il soit réellement porteur suite à un croisement avec la race Holstein qui ne serait pas connu dans le pedigree.

Nous n'avons pas pris en compte la possibilité d'erreur de génotypes dans la construction de nos filtres d'exploitation des données de séquences. Ces deux discordances montrent que nous aurions dû intégrer une notion de qualité de génotypes pour chaque individu considéré. De ce fait, la liste de variants présentée précédemment est non-exhaustive et nous sommes peut-être passés à côté de certaines mutations d'intérêt.

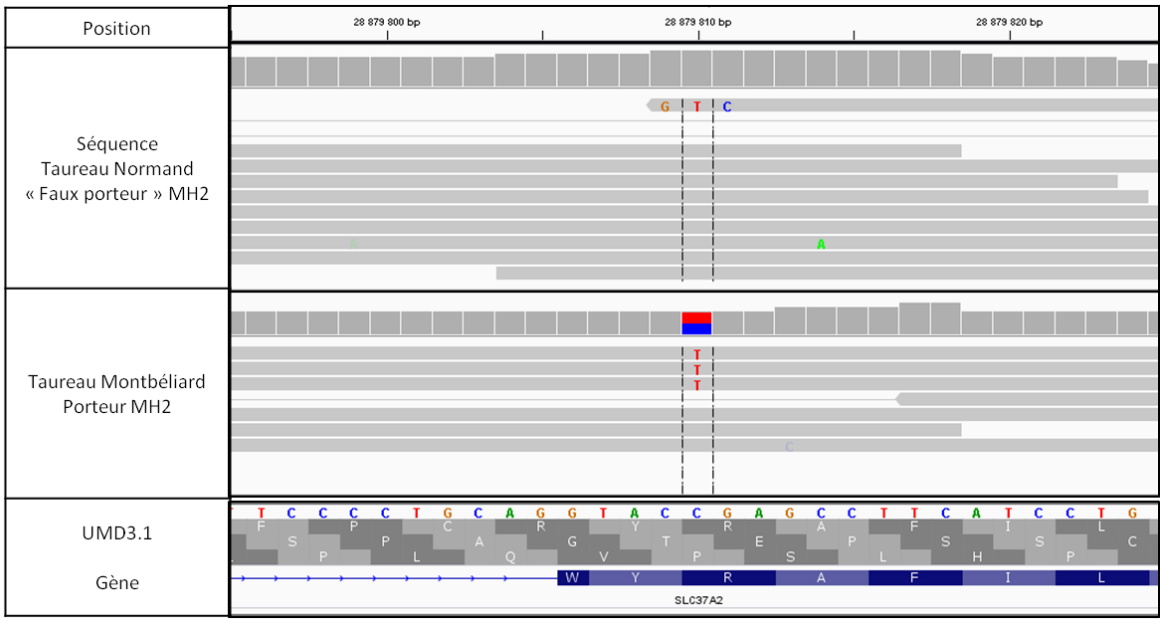

Figure 10. Visualisation IGV (Integrative Genome Viewer) de la mutation candidate SLC37A2 g.chr29:28,879,810 C>T associée à l'haplotype de mortalité embryonnaire MH2.

B. Variants intégrés à la puce EuroG10K et qualité de génotypage

L'objectif principal de cette étude était d'identifier parmi les variants de la séquence ceux potentiellement délétères et de confirmer leur effet sur les animaux présents en ferme. Ceci nécessite d'identifier les animaux homozygotes mutants. Pour cela nous avons intégré des variants d'intérêt, sélectionnés à partir de la liste établie précédemment, sur la puce de génotypage Illumina EuroG10K et réalisé un génotypage à grande échelle sur les animaux analysés chaque année dans le cadre de la sélection génomique française. Cette stratégie permet d'accumuler rapidement une grande quantité de données de typages dans la population tout en ayant accès pour ces individus aux informations de pedigree, de génotypes élaborés (génotypes 50k imputés et phasés) et par la suite de phénotypes lorsqu'ils entreront en production. Cependant nous avons rencontré un certain nombre de difficultés, non envisagées au début du projet qui ont impacté la récupération et l'analyse des données de typage. Nous avons choisi de les exposer dans ce paragraphe qui détaille les SNP retenus et les points techniques critiques rencontrés dans ce projet.

1. Variations candidates sélectionnées pour un génotypage à grande échelle

Nous avons basé nos critères de sélection sur l'annotation fonctionnelle des gènes et les phénotypes observés dans d'autres espèces (voir méthodes). Nous avons retenu 809 variants dont 16% (N=116) génèrent une perte de fonction et 84% (N=684) une substitution délétère d'un acide aminé (

Tableau 7).

Ces variants candidats touchent 657 gènes différents, dont 46% (N=303 gènes, 388 variants) sont associés au moins une fois à un phénotype murin létal (embryonnaire, néonatal ou postnatal) ou de fertilité. Ce sont des variants pour lesquels nous nous attendons en priorité à observer un déficit en homozygotes à partir des résultats de génotypage. Un faible pourcentage d'entre eux a été perdu lors de la confection des puces de génotypage par ILLUMINA. En effet, certaines sondes d'oligonucléotides ne peuvent être synthétisées et incluses sur la puce.

Au final, 3% (N=25) des variants proposés ont été perdus et 784 variants sur les 809 proposés ont réellement été génotypés sur la puce.

Tableau 7: Distribution des variants retenus pour génotypage sur puce à SNP en fonction de leur effet fonctionnel prédit sur la protéine et de la race dans laquelle ils ont été identifiés

	Brune	Charolaise	Holstein	Limousine	Montbéliarde	Normande	Total
Substitution	149	53	176	85	133	88	684
Gain de codon stop	10	4	21	9	1	7	52
Décalage du cadre lecture	2		8	3	9		22
Modification épissage	1	2	28	4	7		42
Modification codon initiateur/codon stop	2	1	1		1		4
InDels avec conservation cadre de lecture	2		1	1	1		5
Total	**166**	**60**	**235**	**102**	**152**	**95**	**809**

2. Elimination des SNP pour échec au génotypage

Un des points critiques de cette étude a été de s'assurer de la fiabilité des génotypes déterminés pour un individu. En nous appuyant principalement sur le *call rate* et les étapes de définition des clusters de génotypes, nous avons identifié au total 239 variants pour lesquels nous considérons que les résultats de génotypage ne sont pas fiables : soit 59 pour lesquels les clusters de génotypes ne peuvent pas être définis et 180 pour lesquels le *call rate* par SNP reste inférieur à 95% (Figure 11).

Parmi ces 180 variants, la majorité est constituée de variants dont les clusters n'ont pas été retravaillés en seconde intention. Si l'on exclut toute erreur de récupération des données de typage, ceci constitue une autre évidence que la définition des clusters de génotypes aurait dû être revue sur l'ensemble des variants. Au final pour chaque race étudiée, cela représente entre 23 et 38% de perte par rapport aux variants sélectionnés au départ.

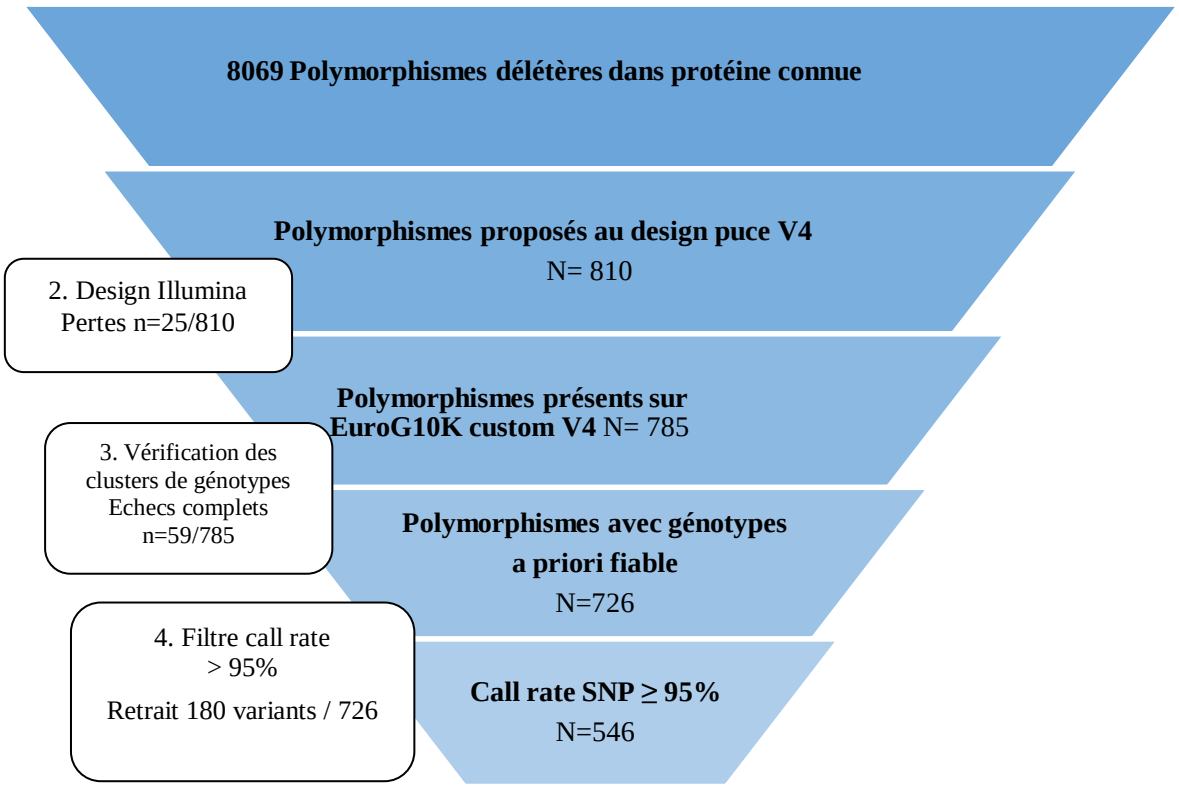

Figure 11: Etapes de sélection et filtres appliqués aux variants après analyse des résultats de génotypage

3. Vérification de la cohérence de l'annotation entre Ensembl et UCSC

Au cours de l'étude des typages, les résultats de quelques variants particuliers nous ont mis sur la piste de variants dont l'annotation réalisée sur la base des transcrits prédits par Ensembl était inexacte. Par exemple, le variant g. chr1:264710G>T, une substitution d'un acide aminé affecte le gène *Chromobox protein 3* (CBX3) candidat intéressant de par son association chez la souris avec un phénotype de mortalité embryonnaire ainsi que des problèmes de fertilité chez le mâle. Il ségrégerait en race Montbéliarde avec une fréquence de 2% et présenterait un déficit significatif en individus homozygotes attendus pour l'allèle alternatif. Cependant, nous avons identifié une incohérence sur la localisation de CBX3 entre les différentes bases de données d'annotation du génome. UCSC et RefSeq (NCBI) placent ce gène sur le chromosome 4 (Chr4:70184737-70195742) alors que Ensembl prédit un transcrit codant sur le chromosome 1 et un pseudo-gène aux positions du chromosome 4. La présence sur le génome de deux copies de ce gène impacte les résultats de génotypage de la variation. L'une des copies portant toujours l'allèle de référence, celui-ci est toujours détecté de sorte que l'on n'observe que des individus homozygotes pour l'allèle de référence et hétérozygotes.

A partir de cet exemple, nous avons choisi de comparer à nouveau toutes les positions testées par génotypage avec les annotations Ensembl et la position du gène sur UCSC. Les 25 variants pour lesquels nous avons identifié ce type d'incohérence ont été écartés.

C. Première analyse des résultats de typage et identification de mutations candidates dans les trois grandes races laitières françaises

L'objectif de cette étude de génétique inverse était de détecter à partir de données de séquence des mutations candidates responsables d'anomalies encore non détectées dans les populations bovines. Nous avons sélectionné des variants sur la base de la fonction du gène qui pourraient avoir un réel effet délétère sur le phénotype des individus porteurs (mortalité embryonnaire, juvénile, animaux anormaux …). Les animaux pour lesquels nous disposons de génotypes sont issus de la population bovine française et sont génotypés dans le cadre de la sélection génomique. Ils constituent donc des individus a priori en bonne santé lors de la réalisation du prélèvement. L'observation d'un déficit traduit des animaux absents ou non sélectionnés pour le génotypage et potentiellement atteints d'une anomalie.

Nous avons donc étudié dans un premier temps la distribution de génotypes sur les différentes populations génotypées pour identifier des mutations candidates dont l'effet fonctionnel pourrait être caractérisé par la suite. Le nombre de typages accumulé étant trop faible dans les races Brune, Limousine et Charolaise, nous avons fait le choix de ne présenter que les résultats sur les variants étudiés dans les races Montbéliarde, Normande et Holstein. Chaque variant a été étudié à partir des résultats de typage de la race dans laquelle il a été sélectionné.

1. Etude générale des fréquences alléliques : un taux élevé de SNP monomorphes

L'allèle alternatif délétère n'a pas été observé pour 32% des variants (N=95/306), soit respectivement 5, 33 et 57 variants monomorphes sur les données de génotypage des races Normande, Montbéliarde et Holstein (Tableau 8). Ces variants monomorphes sont probablement de fausses variations (ou faux-positifs) détectées dans les données de séquence. Elles sont liées à des artéfacts de séquençage ou des erreurs d'alignement des lectures sur le génome de référence (régions répétées ou hautement polymorphes) (Nielsen *et al.*, 2011).

Le taux de faux positifs est également plus élevé lorsque la couverture de séquençage des individus est faible et que la détection est supportée par un seul individu, ce qui est le cas pour la majorité des variants de cette étude. En moyenne, les taux de faux positifs rapportés dans des études similaires se situent entre de 5 et 10% (MacArthur *et al.*, 2012 ; Das *et al.*, 2015 ;). Avec 32% nous avons un enrichissement en fausses variations dans notre sélection de variants.

Pour limiter les faux positifs parmi les variants sélectionnés, nous aurions pu appliquer des points de contrôles supplémentaires. Par exemple nous aurions pu écarter les variations avec un nombre de

lectures supportant l'allèle alternatif trop faible et/ou fixer un nombre minimal de porteurs observés dans le jeu de données. Cependant, le risque sur des couvertures de séquençage faibles est d'écarter des variants rares avec un intérêt potentiel, ce que nous avons cherché à éviter.

Tableau 8: Distribution des variants en fonction de leur fréquence parmi les typages des individus de race Normande, Montbéliarde et Holstein.

	Montbéliarde	**Normande**	**Holstein**
Monomorphes	33 (38%)	5 (7%)	57 (39%)
<1%	12 (14%)	7 (10%)	40 (27%)
1-5%	26 (30%)	30 (42%)	34 (23%)
5-10%	12 (14%)	23 (32%)	12 (8%)
>10%	4 (5%)	6 (8%)	4 (3%)
Total	**87**	**71**	**147**

Pour l'étude de ces variants rares, il est nécessaire de disposer d'un grand nombre d'individus génotypés de façon à distinguer un déficit significatif en homozygotes, si celui-ci existe. Les effectifs disponibles au moment de cette étude ne permettent pas d'apporter une interprétation aux résultats de typage des SNP trop rares. C'est pourquoi nous ne détaillerons pas ici les variants dont la fréquence est inférieure à 1% sur la base des animaux génotypés. Toutefois, ces variants seront étudiés dans le cadre d'études faisant suite à cette thèse, avec l'ajout de nouveaux génotypes, le développement de tests spécifiques sur haplotype et l'imputation des génotypes pour ces mutations sur la population génotypée précédemment avec la puce BovineSNP50.

2. Analyse des polymorphismes sans homozygotes pour l'allèle alternatif

Sur les polymorphismes étudiés avec une fréquence supérieure à 1%, nous observons seulement six polymorphismes sans homozygotes pour l'allèle alternatif délétère dans la race dans laquelle ils ont été choisis (Tableau 9). Parmi eux, seul le SNP chr11 :72399397T>C, ségrégant à une fréquence de 3,5% en race Normande, possède un déficit significatif en homozygotes pour l'allèle alternatif (0 vs 9 ; p-value= $2.8x10^{-3}$). Ce SNP génère une substitution d'un acide aminé dans la protéine multifonctionnelle CAD (p.Y452C). Elle assure chez les animaux trois activités enzymatiques, C*arbamoyl-Phosphate Synthetase 2 (*CSP2*), Aspartate Transcarbamylase (*ATC-ase*)* et *Dihydroorotase* (DHO-ase), qui sont nécessaires à la réalisation des trois premières étapes de la voie de synthèse *de novo* des pyrimidines (Grande-Garcia *et al.*, 2014). Ici, l'acide aminé substitué est situé dans le premier domaine à activité CSP-ase de CAD. Il est totalement conservé chez l'ensemble des vertébrés, ainsi que dans la protéine

mitochondriale paralogue CSP. Ceci suggère une contrainte évolutive forte pour l'activité CPS-ase de l'enzyme (Figure 12).

Tableau 9: Variants potentiellement délétères avec une absence d'homozygotes sur les animaux disposant d'un génotype

Race	Chr	Position	Gène	Ref /Alt	Conséquence	MAF (%)	Typages	OBS (Alt/Alt)	ATT (Alt/Alt)	p-value
NO	19	41557859	*KRT24*	C/T	p.E460K	1,03	5673	0	1	0,8821
	13	62409209	*ASXL1*	C/T	p.T600M	1,14	5687	0	1	0,7680
	14	60995448	*ZFPM2*	G/C	p.P569R	2,44	7572	0	5	0,0519
	11	**72399397**	***CAD***	**T/C**	**p.Y452C**	**3,50**	**7571**	**0**	**9**	**0,0028**
MO	8	105106696	*COL27A1*	C/T	p.R121C	1,07	20746	0	2	0,215
	24	7434109	*RTTN*	C/T	p.S706L	1,11	20744	0	3	0,192

NO : Normande, MO : Montbéliarde, MAF : fréquence de l'allèle alternatif, OBS(Alt/Atl) : nombre d'homozygotes alternatifs observés, ATT(Alt/Alt) : nombre d'homozygotes alternatifs attendus.

Plusieurs mutations délétères dans les gènes des voies de biosynthèse *de novo* des purines et pyrimidines, molécules essentielles dans de plusieurs métabolismes cellulaires, ont déjà été identifiées comme responsables de mortalité embryonnaire chez le bovin (les gènes *GART* et *PFAS* pour les purines et le gène *UMPS* pour les pyrimidines ; Fritz *et al.*, 2013 ; Michot et al., submitted ; cf. chapitre précédent).

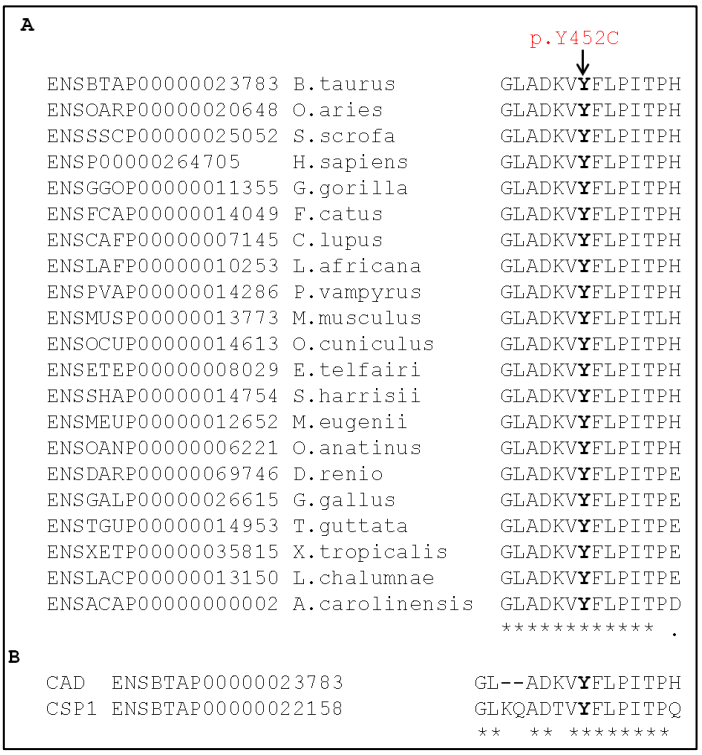

Figure 12: Conservation de l'acide aminé p.Y452 dans le gène CAD.

A. *Alignement des séquences protéiques chez les vertébrés.*
B. Alignement entre CAD et son paralogue CSP1 chez le bovin.

Sur la même voie que *CAD*, nous pouvons souligner le gène codant pour l'uridine monophosphate synthase *(UMPS*) qui intervient dans les deux dernières étapes de la biosynthèse. Une mutation faux-sens de ce gène est responsable du syndrome de déficience de cette enzyme en en race Holstein qui est létal à l'état embryonnaire pour les individus homozygotes mutés (OMIA 000262-9913 ; Schwenger *et al.*, 1993). L'importance de *CAD* a également été récemment démontrée chez l'homme où des mutations de ce gène sont mises en cause dans un syndrome congénital sévère de défauts de la glycosylation (NG *et al.,* 2015). En parallèle, au cours d'une mise à jour de l'étude des haplotypes en déficit en homozygotes, le polymorphisme *CAD* g. chr11:72399397T>C a été identifié comme seule mutation candidate associée à un nouvel haplotype en déficit complet détecté sur le chromosome 11 en race Normande (Fritz S et Capitan A, communication personnelle). Retrouver le même résultat par deux approches complémentaires confirme le polymorphisme dans *CAD* en tant que nouvelle mutation récessive létale ségrégant en race Normande. Le polymorphisme *CAD* g. Chr11 :72399397T>C constitue le seul résultat significatif sur cette première étude.

Cependant, trois autres variants sans homozygotes et avec un léger déficit seraient des candidats potentiels. Ce sont les SNP g.chr8:105106696C>T et g.chr 24:7434109C>T en race Montbéliarde pour lesquels nous n'observons aucun homozygote alternatif contre 3 attendus, ainsi que le SNP g. chr14:60995448G>C étudié en race Normande avec 0 homozygote observé contre 5 homozygotes attendus (p-value=0.0519). Ils sont situés respectivement dans les gènes codant pour la chaîne alpha 1 de la fibre de collagène de type XXVII (COL27A1), la rotatine (RTTN) et la *zinc finger protein* (ZFPM2), impliqués dans des fonctions biologiques importantes et déjà associés à des phénotypes délétères sévères chez l'homme, la souris ou le bovin.

En effet, *COL27A1* intervient au niveau du développement des os et du cartilage et il est associé chez l'homme au syndrome de Steel, une forme récessive d'osteochondrodysplasie (Gonzaga-Jauregui, 2015). *RTTN* appartient à la cascade de gènes régulant la mise en place de l'axe bilatéral lors du développement embryonnaire chez les vertébrés. Il est associé au syndrome humain de polymicrogyrie, une anomalie de développement du cortex cérébral qui se caractérise par une microcéphalie accompagnée d'un retard mental et de développement (OMIM 614833, Kheradmand *et al.*, 2012 ; Shamseldin *et al.*, 2015). Enfin, le gène *ZFPM2* est un facteur de transcription qui régule la morphogenèse cardiaque. Il est associé chez la souris à une mortalité embryonnaire liée à des anomalies du développement du système cardiovasculaire et impliqué chez l'homme dans les syndromes de Tétralogie de Fallot et d'hernies congénitales du diaphragme (OMIM 610187, Tevosian *et al.*, 2000).

Compte tenu des phénotypes sévères associés à ces trois gènes, l'augmentation du nombre de typages et les typages d'individus nés d'accouplements à risques nous permettront de confirmer ou infirmer le déficit en homozygotes supposé.

3. Polymorphismes avec un déficit partiel en homozygotes

Sur les trois races étudiées, 139 variants avec une MAF >1% ont été observés avec au moins 1 homozygote pour l'allèle alternatif. Nous supposons les animaux génotypés en bonne santé (apparente) au moment de la réalisation du prélèvement biologique pour le génotypage. De ce fait, ce résultat suggère que ces polymorphismes n'ont pas ou peu d'effet délétère comparé au phénotype potentiellement attendu. C'est en particulier le cas pour les gènes associés à un phénotype létal chez la souris ou une anomalie génétique sévère chez l'homme ou une autre espèce.

Parmi ces polymorphismes, nous avons identifié 25 variants avec un déficit partiel significatif en homozygotes pour l'allèle alternatif et un écart plus ou moins grand par rapport aux proportions de génotypes attendues (entre 9 et 89 % de différences) (Tableau 10). La détection de ce type de variants a également été rapportée par Charlier *et al.,* (2016). Le déficit partiel suggère que l'allèle alternatif prédit délétère n'est pas létal, mais que les homozygotes sont tout de même soumis à un effet de contre sélection, ce qui en fait des cibles privilégiées d'étude.

L'interprétation que nous pouvons faire de ces polymorphismes reste assez diverse. Malgré les précautions prises pour conserver des variants fiables, il est possible que le déséquilibre provienne d'erreurs de génotypage. Elles entrainent une mauvaise attribution des génotypes des individus, ce qui biaise les fréquences observées. Généralement, le test d'équilibre de HW est utilisé pour écarter les marqueurs avec ce profil, en particulier pour les variants avec une fréquence élevée pour l'allèle alternatif. Dans nos données, ce sont ces variants qui ont en proportion les plus faibles différences observées avec un effet significatif probablement lié aux fréquences élevées (par exemple le variant dans le gène HMCN1 avec une fréquence de 28%).

Nous envisageons également une association forte avec un allèle très délétère proche, responsable par exemple d'une anomalie génétique déjà identifiée ou non. Ce cas de figure pourrait expliquer le déficit observé de 1 homozygote contre 9 attendus pour le variant g.chr21:20898337G>C situé dans le gène MFGE8 (*milk fat globule-EGF factor 8)* ségrégant en race Holstein. Nous avons sélectionné ce variant « à l'aveugle » de par son association avec un phénotype de fertilité chez la souris. Cependant, ce SNP est proche de la délétion g.chr21 :21184870_21188198del située dans le gène *Fanconi Anemia Complementation Group I* (*FANC1* ; p.V877Lfs27X*) responsable de l'anomalie *Brachyspina* (Charlier *et al.*,2012). L'allèle délétère du variant de *MFGE8* est probablement en déséquilibre de liaison avec la délétion du gène *FANCI,* ce qui induit le déficit en individus homozygotes observés (Fritz *et al.*,2013). La même hypothèse pourrait être appliquée au SNP g. chr6 : 68035610C>T situé dans le gène *CORIN* et proche de l'haplotype de mortalité embryonnaire MH5 (chr6 :73,3-74,4 Mb ; Fritz *et al.*, 2013). Dans ce cas, l'association reste à explorer et n'exclut pas un effet délétère du variant.

On peut aussi envisager une association avec un QTL ayant un effet défavorable sur un caractère de production, ce qui entrainerait une élimination des individus homozygotes. Un exemple de ce type serait le variant g. chr14 :9483691 G>A en race Montbéliarde et situé dans le gène de la Thyroglobuline (TG), un précurseur des hormones thyroïdiennes T3 et T4. Ce gène est associé à l'anomalie génétique du gonflement de la glande thyroïdienne (« goitre ») chez le bovin mais également proposé comme gène candidat associé à des QTL de croissance et de conformation dans les races allaitantes (CattleQTLdb : http://www.animalgenome.org ; Ribeca *et al.*, 2013).

Enfin, on peut imaginer que les mutations ont bien un effet délétère mais associé à une pénétrance incomplète ou à une apparition tardive des symptômes, ou tout du moins postérieure au prélèvement pour la sélection génomique. A ce titre, les variants avec moins de 50% d'homozygotes observés par rapport à l'attendu sont des candidats particulièrement intéressants: par exemple les variants *SCL16A4* (p.R245X) et *ABCA4* (p.W273X) en race Normande, ainsi que le gène *FREM1* (p. R1131X) en race Holstein.

Dans le cas des autres variants avec homozygotes, il est possible que la mutation ait un effet sans impact sur la carrière des animaux (par exemple une variation de coloration) ou, là encore, que les symptômes apparaissent plus tardivement dans la vie de l'animal. Leur impact est faible sur la filière, comme nous l'avons décrit pour la mutation de *RP1* induisant une perte progressive de vision en race Normande (voir chapitre 6 ; Michot *et al.*, 2016). Cependant, caractériser ces mutations permettrait d'améliorer l'information sur la fonction des gènes bovins. Pour cela, il sera nécessaire d'établir un examen clinique approfondi des individus homozygotes pour ces mutations. .

Tableau 10: Variants potentiellement délétères avec un déficit partiel en homozygotes.

Race	Chr	Position	Gène	Ref/Alt	Conséquence	MAF(%)	Typages	OBS (Alt/Alt)	THEO (Alt/Alt)	DIFF	p-value
MO	5	28508756	*SLC4A8*	C/T	p.A429V	3,0	20733	7	19	-63%	7,54E-03
	1	20798583	*USP25*	C/T	p.R23Q	3,1	20744	11	20	-46%	4,51E-02
	6	68035610	*CORIN*	C/T	p.D273N	4,4	20729	22	40	-45%	4,13E-03
	24	45847563	*SLC14A2*	G/A	p.V612M	4,9	20742	28	50	-44%	1,51E-03
	13	61925492	*FOXS1*	G/A	p.A28V	5,3	20741	33	58	-43%	8,09E-04
	13	39533047	*SLC24A3*	G/A	p.A56T	3,7	20742	17	28	-39%	4,09E-02
	6	57250328	*ARAP2*	C/T	p.M253I	6,5	20738	55	88	-38%	1,83E-04
	14	9483691	*TG*	G/A	p.A1046V	5,3	16072	31	46	-32%	2,44E-02
	21	47361830	*SLC25A21*	CT/C	Site accepteur épissage	7,0	16064	54	79	-32%	2,79E-03
	23	7328421	*SLC39A7*	G/A	Site donneur épissage	8,1	20710	93	135	-31%	1,13E-04
	28	28008574	*CDH23*	T/G	p.V878G	8,1	16073	81	105	-23%	1,36E-02
	11	31111605	*FSHR*	G/A	p.R484C	16,3	16073	370	425	-13%	1,45E-03
NO	3	33169184	*SLC16A4*	C/T	p.R245X	4,8	7574	7	18	-60%	1,10E-02
	24	25915691	*DSG2*	G/C	p.P990R	4,0	7571	5	12	-60%	4,04E-02
	3	49584965	*ABCA4*	G/A	p.W273X	4,7	7572	8	17	-53%	2,84E-02
	16	75356912	*DIEXF*	C/T	p.G479E	6,3	7574	18	30	-41%	2,18E-02
	28	27685674	*SLC39A3*	C/A	p.R54L	6,3	7572	18	30	-41%	2,48E-02
	2	107907622	*ANKZF1*	C/T	p.R628W	6,1	7568	18	28	-36%	4,86E-02
	3	33324979	*KCNC4*	C/G	p.R13P	11,6	7557	81	103	-21%	1,70E-02
	16	68684721	*HMCN1*	C/T	p.R4732W	28,9	7568	579	635	-9%	1,83E-03
HO	21	20898337	*MFGE8*	G/C	p.R177G	1,7	30959	1	9	-89%	9,00E-03
	15	35646375	*ABCC8*	C/T	p.R1582C	2,0	30962	4	12	-68%	2,17E-02
	8	29433977	*FREM1*	C/T	p.R1131X	2,5	30966	10	19	-48%	2,86E-02
	3	49669134	*ABCA4*	C/G	p.H1836D	5,8	30961	84	105	-20%	3,39E-02
	29	44107876	*SLC22A20*	C/T	p.A429V	14,8	23216	440	506	-13%	6,42E-04

NO : Normande, MO : Montbéliarde, HO : Holstein, MAF : fréquence de l'allèle alternatif, OBS(Alt/Atl) : nombre d'homozygotes alternatifs observés, ATT(Alt/Alt) : nombre d'homozygotes alternatifs attendus.

III. Conclusion

Dans ce chapitre nous avons présenté les premiers résultats obtenus par l'application d'une stratégie de génétique inverse mise en place dans le cadre de ma thèse. Cette stratégie qui s'appuie sur l'exploitation des données de séquence est relativement nouvelle chez le bovin. Son efficacité a déjà été démontrée avec l'identification et la confirmation de sept nouvelles mutations létales à l'état embryonnaire et une mutation affectant la coloration dans les races Jersiaises, Holstein et Blanc-Bleu Belge (Sartelet *et al.*, 2015, Charlier *et al.*,2016).

Bien que notre travail soit inachevé à ce jour, l'exploitation des données de séquence et de génotypage dans les trois grandes races laitières françaises a mis en évidence une mutation dans le gène *CAD* (g. 11 : 72399397T>C ; p.Y452C) probablement responsable de mortalités embryonnaires en race Normande. Nous rapportons trois autres mutations qui pourraient être létales mais dont la fréquence trop faible sur les données disponibles empêche la validation. Nous identifions aussi plusieurs polymorphismes d'intérêt présentant des déficits partiels, en particulier dans les gènes *FREM1, SLC16A4 et ABCA4*, qui seront à étudier en priorité. Cette étude fournit également des informations de typages pour un grand nombre de polymorphismes qui peuvent être utiles à d'autres projets menés dans l'équipe. En particulier, une partie des variants génotypés sur puce ont été valorisés dans un article de caractérisation des données de séquençage françaises (voir Boussaha *et al.*,2016).

Dans l'application de l'approche de génétique inverse, nous envisagions au départ que la principale difficulté serait la caractérisation phénotypique de certaines mutations chez les animaux, impliquant d'identifier des animaux homozygotes ou issus d'accouplements à risque sur le terrain. C'est au final le développement du génotypage de variants rares sur puce à SNP qui nous a posé le plus de problèmes et entrainé un retard dans la réalisation du projet. Cependant, ces contretemps ont montré la nécessité de s'assurer d'une bonne définition des clusters pour chaque polymorphisme et ont permis de développer des procédures avec le laboratoire de génotypage pour le traitement des polymorphismes inclus sur les puces de typage dans le cadre de la recherche. Dans l'idéal, il faudrait pouvoir retravailler tous les polymorphismes sur un groupe fermé et suffisamment grand d'individus à chaque nouvelle version de puce, puis ré-analyser l'ensemble des individus génotypés. Un autre point concerne le nombre élevé de variants monomorphes au typage et les erreurs d'annotation des polymorphismes qui ont limité l'efficacité de notre approche. Ce sont deux points qu'il faudra améliorer au niveau de la construction des filtres pour l'exploitation des données de séquence.

La caractérisation phénotypique et fonctionnelle des quelques variants candidats identifiés, et plus particulièrement du variant candidat létal dans le gène *CAD*, n'a pas pu être menée à bien au cours de ma thèse. Elle sera poursuivie dans les prochains mois pour caractériser un effet s'il existe. En parallèle, il reste à étudier les typages pour les variants potentiellement délétères dans les trois autres races

analysées (Brune, Limousine et Charolaise) et les variants trop peu fréquents en races laitières, pour lesquelles suffisamment d'information devrait avoir été accumulée pour permettre d'effectuer un tri similaire et repérer des variations avec un intérêt potentiel.

CHAPITRE 6 : APPROCHE DE GENETIQUE INVERSE ET ETUDE DES VARIANTS DELETERES FREQUENTS DANS LES RACES BOVINES EUROPEENNES

Au chapitre précédent, nous nous sommes intéressés aux variants délétères race-spécifiques dont la fréquence reste généralement faible dans les populations bovines. Cette approche permet d'identifier des variations létales ou avec un effet très négatif sur le phénotype des animaux avant que l'anomalie ne soit identifiée. Cependant elles ne représentent qu'une petite partie de l'ensemble des variants potentiellement délétères prédits dans le génome bovin. Dans la prochaine étude, nous avons choisi d'explorer les conséquences de variants délétères plus fréquents identifiés à partir des données de séquence.

Article 3 : Identification d'une mutation ancienne dans le gène *RP1* responsable d'une dégénérescence progressive de la rétine chez le bovin

Michot P, Chahory S, Marete A, Grohs C, Dagios D, Donzel E, Aboukadiri A, Deloche MC, Allais-Bonnet A, Chambrial M, Barbey S, Genestout L, Boussaha M, Danchin-Burge C, Fritz S, Boichard D and Capitan A.

A reverse genetic approach identifies an ancestral frameshift mutation in *RP1* causing recessive progressive retinal degeneration in European cattle breeds

Michot *et al.*, *Genet Sel Evol* (2016) 48:56

Michot *et al. Genet Sel Evol (2016) 48:56*
DOI 10.1186/s12711-016-0232-y

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

A reverse genetic approach identifies an ancestral frameshift mutation in *RP1* causing recessive progressive retinal degeneration in European cattle breeds

Pauline Michot[1,2], Sabine Chahory[3], Andrew Marete[1,4], Cécile Grohs[1], Dimitri Dagios[3], Elise Donzel[3], Abdelhak Aboukadiri[1], Marie-Christine Deloche[1,2], Aurélie Allais-Bonnet[2,5], Matthieu Chambrial[6], Sarah Barbey[7], Lucie Genestout[8], Mekki Boussaha[1], Coralie Danchin-Burge[9], Sébastien Fritz[1,2], Didier Boichard[1] and Aurélien Capitan[1,2*]

Abstract

Background: Domestication and artificial selection have resulted in strong genetic drift, relaxation of purifying selection and accumulation of deleterious mutations. As a consequence, bovine breeds experience regular outbreaks of recessive genetic defects which might represent only the tip of the iceberg since their detection depends on the observation of affected animals with distinctive symptoms. Thus, recessive mutations resulting in embryonic mortality or in non-specific symptoms are likely to be missed. The increasing availability of whole-genome sequences has opened new research avenues such as reverse genetics for their investigation. Our aim was to characterize the genetic load of 15 European breeds using data from the 1000 bull genomes consortium and prove that widespread harmful mutations remain to be detected.

Results: We listed 2489 putative deleterious variants (in 1923 genes) segregating at a minimal frequency of 5 % in at least one of the breeds studied. Gene enrichment analysis showed major enrichment for genes related to nervous, visual and auditory systems, and moderate enrichment for genes related to cardiovascular and musculoskeletal systems. For verification purposes, we investigated the phenotypic consequences of a frameshift variant in the *retinitis pigmentosa-1* gene segregating in several breeds and at a high frequency (27 %) in Normande cattle. As described in certain human patients, clinical and histological examination revealed that this mutation causes progressive degeneration of photoreceptors leading to complete blindness in homozygotes. We established that the deleterious allele was even more frequent in the Normande breed before 1975 (>40 %) and has been progressively counter-selected likely because of its associated negative effect on udder morphology. Finally, using identity-by-descent analysis we demonstrated that this mutation resulted from a unique ancestral event that dates back to ~2800 to 4000 years.

Conclusions: We provide a list of mutations that likely represent a substantial part of the genetic load of domestication in European cattle. We demonstrate that they accumulated non-randomly and that genes related to cognition and sensory functions are particularly affected. Finally, we describe an ancestral deleterious variant segregating in different breeds causing progressive retinal degeneration and irreversible blindness in adult animals.

*Correspondence: aurelien.capitan@jouy.inra.fr
[1] UMR 1313 GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
Full list of author information is available at the end of the article

Background

Domestication has had a dramatic effect on the genomes of plant and animal species. Reduction of environmental pressure combined with rapid growth of populations after strong demographic bottlenecks have resulted in relaxation of purifying selection and accumulation of deleterious mutations [1–5]. In the last 150 years, this phenomenon termed "the cost of domestication" has been particularly amplified in cattle because of the creation of breeds from a limited number of founder animals, overuse of a few elite sires with artificial insemination (AI) and intensive selection on specific traits. As a consequence most bovine breeds experience regular outbreaks of recessive genetic defects. With the advent of high-throughput genotyping and next-generation sequencing, efficient methods have been developed to identify the underlying mutations in record time and with a limited number of available cases [6, 7]. However, such approaches rely on the observation of affected animals with distinctive symptoms. It can be anticipated that the genetic defects reported so far represent only the tip of the iceberg and that many recessive mutations resulting in embryonic mortality or in non-specific symptoms, which can be confounded with those of common diseases, remain to be discovered. In addition to the influence of genetic drift and hitch-hiking, the frequency of some deleterious mutations, which would be detrimental in the wild, may have been involuntarily increased by artificial selection on behavior, coat color, morphological or production traits. This is the case for example for double-muscling, which causes dystocia [8, 9], and for a series of mutations under balancing selection [10, 11].

The increasing number of available whole-genome sequences (WGS) has recently opened new research avenues such as reverse genetics to investigate recessive defects. This strategy seems particularly suitable in cattle for which the sequencing of the most influential AI bulls of each breed (e.g. 1000 bull genomes project [12]) enables the identification of the vast majority of the non-private deleterious mutations that segregate in these populations. Furthermore, the inclusion of a subset of these polymorphisms into single nucleotide polymorphism (SNP) chips that are used for genomic selection should facilitate the detection of homozygotes (or of a deficit in homozygotes) for deleterious alleles among the tens of thousands of animals genotyped each year. In parallel, crossing genotyping data with pedigree information should enable the detection of severely affected homozygotes among animals that are born from at risk matings, which would not have been genotyped for genomic selection purposes. Finally, bovine populations provide an important number of cases available for sampling and experimental study to evaluate the functional consequences of the mutation, which is hardly possible in humans.

The purpose of this study was twofold: (i) to characterize the genetic load of 15 beef and dairy breeds using whole-genome sequencing data from the 1000 bull genomes consortium [12] and (ii) to prove that widespread harmful mutations remain to be detected in our cattle populations by characterizing the effect of a frameshift mutation in the *retinitis pigmentosa-1* (*RP1*) gene, which segregates in Normande cattle and other European breeds.

Methods

Ethical statement

Blood and ear biopsies were collected by veterinarians or by agricultural technicians licensed by the French Departmental Breeding Establishments [Etablissements Départementaux de l'Elevage (EDE)] during routine ear tagging, sampling for annual prophylaxis, paternity testing and genotyping for genetic defects or genomic selection. Ophthalmologic examinations and electroretinograms were approved after ethical evaluation by the ComERC committee (Ethical Committee for Clinical Research at the French Veterinary School of Maisons Alfort (ENVA) (Saisine n°14-01-2015) and performed under sedation controlled by a veterinarian specialized in cattle.

Invasive procedures were performed post-mortem after slaughter for meat production. Experiments reported in this work comply with the ethical guidelines of the French National Institute for Agricultural Research (INRA). All the samples and data analyzed were obtained with the permission of breeders, breeding organizations and research group providers.

Animals

Details on animals used for each analysis are presented in Additional file 1: Table S1.

Filtering of variants from whole-genome sequence data and prediction of their phenotypic consequences

Variants were selected from whole-genome sequence data of 1147 bulls from the 1000 bull genome project (for details on variant calling see Daetwyler et al. [12]). Briefly, raw reads were filtered and trimmed on chastity and quality score, then aligned on the UMD3.1 bovine reference sequence assembly using BWA [13]. SNPs and InDel were called from pooled bam files using SAMtools 0.1.18 mpileup [14]. Variants were then annotated using Ensembl Variant Effect Predictor [15]. Frequencies and allele counts were calculated across and within breeds using vcftools "freq" and "count" options [16]. Filtering consisted in selecting biallelic variants which (i) were

predicted to cause a loss of protein function (i.e. affecting initiator codons, splice acceptor or donor sites, or causing a frameshift, a stop loss or gain, or a missense with a SIFT score of 0 [17]), (ii) had a calling quality (QUAL) above 30, (iii) presented a mapping quality (MQ) score of 59 or 60, (iv) had less than 5 % of animals with missing genotypes, and (v) had a minor allele frequency (MAF) higher than 5 % for at least one breed with a minimum of 20 individuals in the dataset (which means that alleles observed only once were not considered). It should be noted that variants with a SIFT score less than 0.05 are generally considered deleterious. In this study, we chose to retain only missense variants with a SIFT score of 0 to reduce possible artifacts. Furthermore, including missense deleterious variants with a SIFT score between 0.01 and 0.05 would have resulted in considering approximately one fourth of the total number of bovine genes, thus preventing subsequent gene enrichment analysis. In addition, each variant was manually checked to eliminate artifacts due to (i) adjacent substitutions within the same codon which are not accounted for in variant annotation, (ii) errors of annotations after comparing gene annotations from the UCSC and Ensembl genome browsers (http://genome.ucsc.edu, http://www.ensembl.org) annotations, (iii) repeated sequences (downloaded at http://genome.ucsc.edu, accession 21/10/2015). Only variants with a known official gene symbol were considered in the subsequent analyses.

To anticipate the phenotypic consequences of the mutations, annotations were completed by information on genetic syndromes associated with mutations within the same genes in humans (Online Mendelian Inheritance in Man, OMIM; http://www.omim.org) and mouse (Mammalian Phenotypes; http://www.informatics.jax.org) (see Additional file 2: Table S2).

Gene set enrichment analysis

Gene enrichment analysis was performed using Ingenuity Pathway Analysis software (http://www.ingenuity.com/products/ipa/, [18]). We focused on "top canonical pathways" with a p value lower than 0.01 and "diseases and bio functions" annotations with a p value lower than 0.05. Annotations related to cancer and the general pathways entitled "skin lesion" and "liver lesion" were not considered since their results suffer from a bias. Pathways related to drug metabolism, which were not relevant for this study, were also eliminated. In addition, a unique keyword was assigned to each significantly enriched function annotation, with particular attention paid to the attribution of keywords related to subcellular portions, cell types and organs rather than to general processes. When possible, keywords appearing only once were regrouped with higher order items (e.g. cell type changed for organ, or process changed for the category defined by IPA) or with the predefined IPA "categories". Frequency of keywords was used to set the size of the words in the word cloud representation.

Ocular examination

Twenty-three pure and crossbred Normande cows from the INRA experimental facility of Le Pin-au-Haras (Normandy, France) with genotypes available for the frameshift mutation in the *retinitis pigmentosa-1* (*RP1*) gene were selected for ocular examination (see Additional file 1: Table S1). These consisted in four homozygous mutants, nine heterozygous and ten homozygous wild type animals. All these animals were in good health conditions and with no signs of systemic disease at the time of the study. Genotypes were not disclosed to the veterinarian to exclude any bias of personal interpretation. Examinations were performed indoors under ambient light. Visual performance was evaluated by the menace response test and dazzle and pupillary light reflexes (direct and indirect) were assessed with a Finoff transilluminator. Slit-lamp biomicroscopy (Kowa SL-15, Kowa Company) was performed before and after pupillary dilation using one drop of 1 % tropicamide. Fundi were examined by indirect ophthalmoscopy (Heine Omega 100, Heine Optotechnik, GmbH & CoKG) with 28-D and 20-D lenses.

Electroretinogram tests

Electroretinogram tests were performed on two 5.5-years old cows, one homozygous wild type and one homozygous mutant, with a Retiport (Roland Consult, Brandenburg, Germany), under sedation (Xylazine 0.04 mg/kg IM) and after pupillary dilation (tropicamide eyedrops) and blocking of the auriculopalpebral nerve by subcutaneous injection of lidocaine. Topical tetracaïne eyedrops were used to anesthetize the ocular surfaces and corneas were lubricated by topical application of sodium hyaluronate 1.2 % during the test. The following responses were recorded: rod response before and after dark adaptation for 20 min, following a dim white stimulus (0.02–0.03 cd/m^2/s), mixed response following four bright white flashes (2–3 cd/m^2/s) at a rate of 0.1 Hz and cone response following four bright white flashes (2–3 cd/m^2/s) at a rate of 5 Hz.

Genotyping of Normande cattle that were reported to the French National Observatory of Bovine genetic abnormalities for progressive loss of vision

Twenty-eight Normande cows that were reported to the French National Observatory of Bovine genetic Abnormalities (ONAB) with signs of progressive loss of vision and blindness were genotyped for the *RP1* frameshift

mutation. Genomic DNA was extracted from blood or ear biopsies using a standard phenol–chloroform protocol and genotyped by PCR and Sanger sequencing for the Chr14 g.23995411_23995412insA mutation. PCR primers were designed from the UMD3.1 bovine genome assembly with Primer3 software [19]) to span the insertion (left: TGCACAGGAAACCATATTGC and right: TTGCCCTAGTTGTGACATGC). Reactions were performed using the Go-Taq Flexi DNA Polymerase (Promega) according to the manufacturer's instructions on a Mastercycler pro thermocycler (Eppendorf). The resulting amplicons were purified and bidirectionally sequenced by Eurofins MWG (Germany) using conventional Sanger sequencing. Polymorphisms were detected with the novoSNP software [20].

Estimation of the allelic frequency of the *RP1* frameshift mutation

The *RP1* frameshift mutation was included in the Illumina EuroG10K custom SNP chip, which is routinely used for genomic selection in France. Thus, in addition to the 1000 genome dataset, genotypes for this mutation were available for 53,279 Holstein, 40,548 Montbéliarde, 12,106 Normande, 1634 Abondance, 1005 Red Pied Lowland, 698 Tarentaise, 579 Simmental, 507 Vosgienne, and 296 Brown Swiss animals.

Post-mortem ocular examination and histological analysis

The eyes of two homozygous mutant cull cows (aged 8 years) that displayed a severe phenotype and two control cows (one 6-year-old heterozygous Normande and one 8-year-old homozygous wild-type Holstein) were collected post-mortem at the slaughterhouse (SVA Trémorel, France). One eye was dissected on site to perform a visual examination of the eye's fundus and to collect the retina and choroid. Samples were immediately frozen in liquid nitrogen and stored at −80 °C until DNA extraction. The second eye was injected with 2.5 ml formaldehyde and fixed by a 24-h incubation in the same solution. The retina and choroid were subsequently dehydrated in a graded ethanol series, cleared with xylene and embedded in paraffin. Microtome sections. (5 µm, Leica RM2245) were stained with haematoxylin, eosin and saffron (HES). Digital images were obtained with the NanoZoomer 2.0- HT slide scanner (Hamamatzu).

Association with recorded traits

At the time of the analysis, Illumina EuroG10K SNP chip genotype data were available for 7439 Normande animals, which had their sire genotyped with the Illumina BovineSNP50 chip. The Illumina EuroG10K SNP chip comprises the *RP1* frameshift mutation as well as more than 10,000 common SNPs with the Illumina BovineSNP50 chip. Using Fimpute [21] we were able to attribute a genotype for the *RP1* polymorphism to 48,715 additional animals that were previously genotyped with the BovineSNP50 chip. The complete dataset comprised 11,986 Normande cows with phenotype information on three coat colour phenotypes (proportion of white areas on the body; proportion of white areas on the face, and brindling intensity) and on 28 traits that are routinely recorded for genetic evaluations (milk yield, fat content, protein content, fat percent, protein percent, cell score, clinical mastitis, milking speed, stature, chest width, body depth, width at pin bone, rump angle, rear legs side view, rear legs rear view, back muscle, fillet muscle, rear muscle, fore udder attachment, rear udder height, udder balance, teat orientation, front teat distance, udder support, udder depth development, interval between calving and first insemination, fertility at insemination of lactating cows, fertility at insemination of heifers). Associations between the *RP1* frameshift polymorphism and traits were tested using GCTA [22]. Phenotypes were adjusted for environmental effects which were estimated in the national genetic evaluation procedure and assumed to reflect the genetic effect of the animal and a random residual effect. Therefore, the analysis model included only an overall mean, a polygenic effect, the effect of the genotype at the *RP1* frameshift polymorphism, and a residual. The polygenic effect was estimated by using a genomic relationship matrix that was derived from 43,801 SNPs on the Illumina BovineSNP50 chip. Finally, a Bonferroni correction that consisted in dividing the p value by the total number of tests performed was applied to account for multiple-testing.

Across-breed identity-by-descent analysis around the *RP1* frameshift mutation

Identity-by-descent (IBD) analysis was performed to (i) test for the existence of one versus multiple mutation events in the different breeds and (ii) estimate the date of the origin of the mutation(s). For that purpose, phased genotypes for a 1.3-Mb region (Chr14:23474270-24643266; corresponding to the smallest IBD homozygous region detected in the genome of one homozygous mutant Normande AI bull, named Diametre (FR5388012666) were extracted for 35 heterozygous and three homozygous carrier animals identified among the 1147 animals from run 4 of the 1000 bull genomes project. Phasing was performed within the framework of the 1000 bull genome project using BEAGLE [12, 23].

Within this homozygous region of Diametre's genome, 9448 SNPs with the highest quality score (QUAL = 999) were selected and considered as reference haplotypes. For each animal, the rate of homozygous genotypes in opposition with the chosen reference genotypes was calculated

for sliding windows of 100 SNPs. Then, the number of individuals that had at least 5 % of inconsistencies with the reference haplotype was counted and attributed to the position of the 51th SNP in each window. This level of 5 % of inconsistencies was chosen to account for the low sequence coverage of certain animals and for the putative occurrence of de novo mutations over time in the vicinity of the old frameshift mutation. The IBD block around the frameshift mutation was finally defined by windows for which none of the carriers displayed 5 % or more of inconsistencies with the haplotypes of Diametre. For control purposes, the same process was applied to a set of 38 non-carrier animals that were randomly selected among individuals belonging to the same breeds as the carriers.

Estimation of the age of the *RP1* frameshift mutation according to the size of the IBD segment shared among breeds

We considered that two animals that shared an IBD segment of size c (c being the size in Morgan) inherited this segment from a common ancestor that lived 1/(2 c) generations ago. We assumed that, on average, 1 cM corresponds to 1,000,000 bp and that generation intervals range from 5 to 7 years, depending on the breeding system (natural mating population or modern breeding schemes).

Analysis of the changes in frequency of the *RP1* frameshift mutation in the Normande breed

To study the changes in allelic frequency of the *RP1* frameshift mutation in the Normande breed, first we developed a haplotype test using 15,515 animals (1077 homozygous carrier, 6363 heterozygous and 8075 homozygous wild type animals) that were genotyped for this variant with the Illumina EuroG10K custom SNP chip and had been phased and imputed for the Illumina BovineSNP50 markers within the framework of the French genomic selection [24]. The haplotype was fixed to 50 SNPs between SNPs ARS-BFGL-BAC-12159 (Chr14 position 22587081 bp) and ARS-BFGL-NGS-36089 (Chr14 position 25698286). We identified 691 haplotypes among which 12.45 % were associated to the frameshift mutation, 83.79 % were not associated with it and 3.76 % were classified as undetermined (i.e. detected in both homozygous carriers and non-carriers). When applied to all the Normande cattle phased Illumina BovineSNP50 genotyped data, 97.3 % of the haplotypes were assigned a status (27.26 % were associated to the frameshift mutation, 83.79 % were associated to the wild type allele, 1.03 % undetermined) and 2.47 % were classified as not documented due to lack of haplotype information among the animals genotyped with the EuroG10K chip. From these haplotype-allele associations, we estimated the genotypes for 1375 phased Normand AI bulls (born between 1975 and 2015) for the *RP1* frameshift polymorphism. Allelic frequencies were calculated over time for sliding windows of 7 years (i.e. on average one generation) after removing haplotypes without information.

Results and discussion

During domestication, deleterious mutations have accumulated in non random sets of genes

A series of filters was applied to draw a list of non-rare putative deleterious polymorphisms in the most important cattle breeds and to reduce as much as possible the false discovery rate (see "Methods" section). Since this study focused on non-rare variants, putative deleterious polymorphisms with a frequency lower than 5 % in all breeds were not investigated. This analysis yielded 2489 putative deleterious variants (stop lost and gained, frameshift, splice acceptor and donor sites, initiator codon variants and missense variants predicted as deleterious with a score of 0 by SIFT) that segregated at a frequency of 5 % or more in at least one of the 15 breeds represented by at least 20 genomes in run 4 of the 1000 bull genomes project [12] (Fig. 1; for details see "Methods" section). The distribution of these variants was similar in terms of number and type of mutations between breeds in spite of quite different numbers of sequenced animals. This result can be explained by the rather high variant frequencies considered. Interestingly, 89 % (2216/2489) of these polymorphisms were observed in more than one breed and as much as 12 % (308/2489) in all 15 breeds, which indicates (subject to any unregistered crossbreeding event) that the majority of the retained variants existed prior to the splitting of the different cattle populations studied (i.e. at least 500 years ago [25]).

A total of 1923 genes carried a deleterious mutation of which 566 counted two or more. A screening of phenotype databases revealed that 908 genes (1144 variants) were associated to at least one mammalian phenotype in laboratory animals (MGI database) and 375 (corresponding to 395 variants) with an inherited syndrome in humans (OMIM database). From our own interpretation, almost two-thirds of these syndromes described in mouse and humans presented a phenotype that would have been difficult to detect by the different national observatories for genetic defects in cattle (i.e. those affecting metabolism, immunity, cognition) (see Additional file 2: Table S2).

In this selection, we also retrieved five variants that were previously reported to cause major phenotypes in cattle. These comprise mutations that have been favored by artificial selection (i.e. p.Q204X mutation in *MSTN* for double muscling in Charolais [26]), or with a severe but

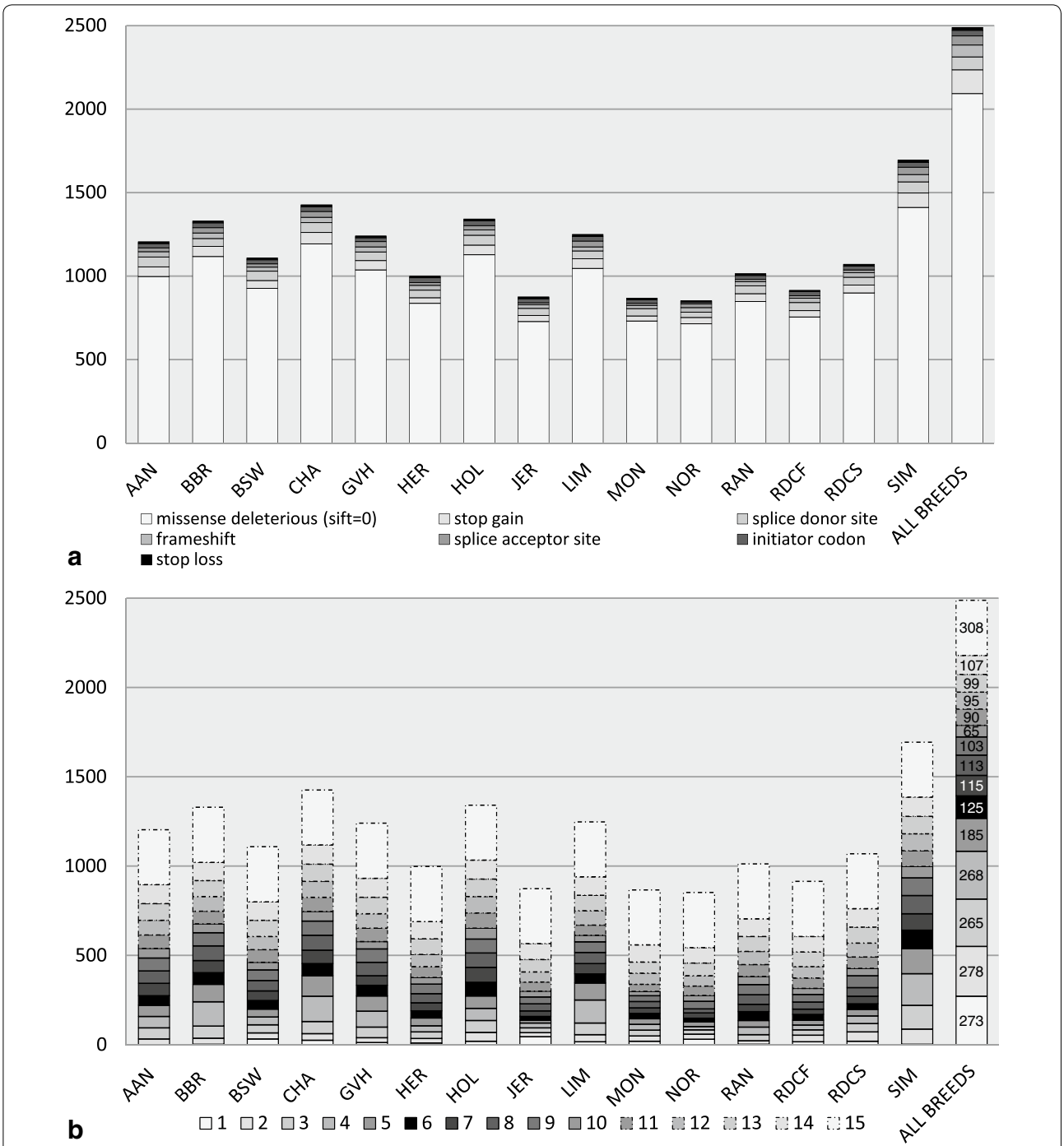

Fig. 1 Details on non-rare putative deleterious variants selected in 15 breeds from the 1000 bull genomes run4 dataset. **a** Distribution of the variants by breed and type of mutations. **b** Distribution for each breed of the number of variants shared with other breeds. Note that only 18.3 % (273/2489) of non-rare putative deleterious variants are breed-specific. *AAN* Aberdeen-Angus, *BBR* Beef Booster Composite, *BSW* Brown Swiss, *CHA* charolais, *GVH* Gelbvieh, *HER* Hereford, *HOL* Holstein, *JER* Jersey, *LIM* Limousine, *MON* Montbéliarde, *NOR* Normande, *RAN* Red Angus, *RDCF* Finnish Red, *RDCS* Swedish Red, *SIM* Simmental

invisible phenotype (i.e. p.R12X and p.R55X nonsense mutations in *SCL37A2* and *CWC15* for embryonic mortality in Montbeliarde [27] and Jersey [28], respectively), or with a mild phenotype that is present in several breeds (i.e. p.R238X mutation in *FMO3* for trimethylaminuria or "fishy-off flavor" of milk [29] and p.W80X mutation in *BCO2* for the "yellow color" of milk and fat [30]). These examples validate that such variants which are deleterious to the protein function may exist and segregate at moderate to high frequencies in cattle breeds.

To obtain an overall picture of the developmental pathways that are affected by our set of variants, we performed a gene enrichment analysis using the ingenuity pathway analysis (IPA) software [18]. This revealed an important enrichment for genes related to nervous system development and function and moderate enrichments for a limited number of other diseases, physiological and biological annotations (see Additional file 3: Tables S3, S4, and S5). We then analyzed the frequency of the keywords that were assigned to each annotation to gain further insight into the organs, tissues or systems represented (Fig. 2) and Additional file 3: Table S6. With 41.5 % of the word counts, the largest cluster was by far composed of words related to nervous, visual and auditory systems, which comprised genes involved in sensorial functions and/or cognition. Indeed, we noted as much as 17.7 % (72/407) of genes related to retina development and function, as well as genes involved in other defects of eye development such as cataract and microphtalmia, and genes associated with deafness (e.g. genes coding for cochlin, *COCH*; otogelin, *OTOG*; otogelin-like, *OTGL*; myosin heavy chain 15, *MYO15A*; and stereocilin, *STRC*) [31–35]. Note that we also detected a number of deleterious mutations in olfactory receptor genes which are not considered by IPA and thus were not accounted for in our analysis.

In addition, we retrieved important genes for neurocognitive functions which are associated with behavioral disorders in humans such as mental retardation, schizophrenia, bipolar disorder or autism. Among other examples, we can cite genes coding for glutamate receptors (*GRIK2* and *GRM7*), glutamate being the most important neurotransmitter in the brain, semaphorins which are involved, among other functions, in axon guidance (*SEMA3A*, *SEMA3B*, *SEMA4A*, *SEMA4D* and *SEMA5A*), calcium voltage channel subunits (*CACNA1C* and *CACNB2*), a receptor for neuroregulin 1 (*ERBB4*), the neurexin-3-alpha protein which has an important role in neuronal function (*NRXN3*), and a post synaptic protein (*SYNGAP1*) [36–45]. Interestingly, only four of the 407 genes from this cluster co-localized with selective sweeps in cattle, chicken, rabbit and/or pig (i.e. *GRIK2* in rabbit, *SEMA3* in pig and chicken, and *ERBB4* and *CACNA1C* in cattle) [46–49].

Therefore, whereas genes that are involved in sensory functions and cognition represent obvious targets of domestication [5, 46], it is unlikely that the variants reported here were positively selected during domestication or subsequent selection processes. More likely, our results indicate that, in a domestic context, such mutations were more tolerated than mutations affecting other systems which are of primary importance for production, reproduction and survival.

Analyzing the frequency of the keywords that were assigned to each IPA annotation also revealed two additional clusters related to cardiovascular (12.8 %), and muscle and skeletal systems (12.3 %), which might be associated with positive selection. These two clusters comprised genes that are associated with selective sweeps and/or production traits such as *MSTN* for double-muscling, *CCNL1* for reduced birth weight, *THADA* for body weight variation, *GOLGA4* for stature, and

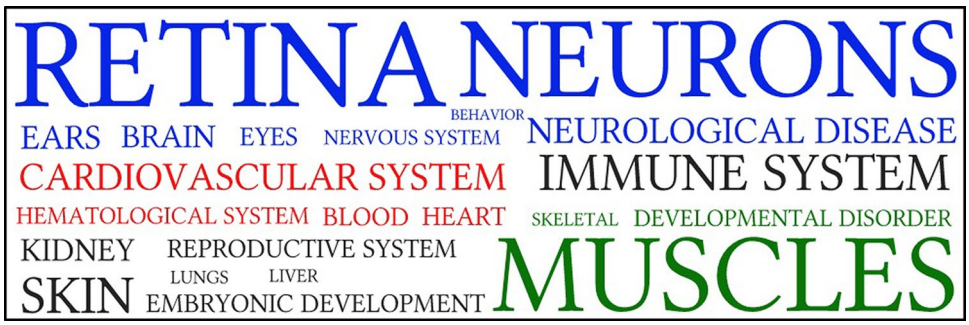

Fig. 2 Word cloud representing the frequency of key-words assigned to significant ingenuity pathway analysis annotations. Only IPA annotations for "top diseases and bio functions" with a p value <0.01 were considered. The size of the font used is proportional to the frequency of each keyword associated with functional annotations. Keywords were clustered into overall related systems: (i) *blue* nervous, visual and auditory systems; (ii) *red* cardiovascular system; and (iii) *green* musculoskeletal systems. They represented respectively 41.5, 12.8 and 12.3 % of the functional annotations considered in the analysis. Annotations related to the two most frequent words, retina and neurons represent, respectively 10.8 and 8.7 % of the total number of annotations

LCORL for stature and skeletal development in cattle [49–51] and *FANCA* for the cardiac system and *NR3C2* for blood pressure in horse [5]. Note that neuro-muscular interactions were also underlined by two IPA canonical pathways (p value <0.01), i.e. the "Agrin interactions at neuromuscular junctions" and "nNOS signaling in skeletal muscles", which are both involved in neuro-muscular diseases and muscle dystrophies. Finally, two other canonical pathways, the "complement system" (part of the innate immune system of an organism) pathways and the "t-RNA charging" (regrouping the key enzymes of protein translation mechanisms in mitochondria) are relevant because of their involvement in two functions that are subject to important selection pressure i.e. immunity and energetic metabolism via mitochondrial function [52, 53] (see Additional file 3: Table S7).

A frameshift mutation in the *RP1* gene causes progressive blindness in cattle

A good candidate for phenotype characterization

For validation, we decided to evaluate the phenotypic consequences of one mutation which was both (i) observed in numerous breeds and (ii) predicted to affect the organ that was most represented in the previous analyses, i.e. the retina. We selected a one base pair insertion (Chr14: g.23995411_23995412insA) that affects the *retinitis pigmentosa-1* gene (*RP1*) which segregates at a particularly high frequency in Normande dairy cattle (Table 1). This mutation is predicted to cause a frameshift at codon 791 and to terminate the protein 13 amino acids later (p. R791KfsX13). If synthesized, the resulting protein would be truncated at 40 % of its normal length and consequently lack two-thirds of its *C*-terminal end.

In humans and mouse, similar truncation mutations in the *RP1* gene, which encodes a microtubule associated protein that is essential for the organization of the outer segments of the photoreceptors in the retina, have been reported to cause autosomal dominant and recessive retinitis pigmentosa [54–57]. Retinitis pigmentosa is a form of inherited degenerative retinal disorder that is characterized by progressive death of photoreceptor cells. Symptoms typically start with loss of night vision due to degeneration of rod-photoreceptors, followed by degeneration of cone-photoreceptors leading to loss of central vision and eventually to complete blindness [58].

For decades, Normande cattle have been considered to have poor eyesight, with older animals showing a typical loss of night vision or blindness. Because it was considered as a breed-specific trait, only a few cases had been reported to the French National Observatory of Bovine genetic Abnormalities (ONAB) and no genetic studies had been initiated. In a first attempt, we genotyped by PCR and Sanger sequencing 28 Normande cows that had been declared to the ONAB for partial or total blindness with no other indication of external eye affection. We observed a significant increase in the number of homozygous mutants (Chi square; p value = 0.003) in this group compared to the population of sequenced Normande founder sires (Table 2), which suggests that this frameshift mutation is responsible for a non-negligible part of the loss-of-vision problems observed in Normande cattle. As a consequence, we decided to include this variant in the EuroG10K SNP chip, to collect genotype information on the French bovine population that is genotyped for genomic selection and to identify carriers for subsequent phenotype characterization.

Table 1 Frequency of the *RP1* frameshift allele among breeds represented in run4 of the 1000 bull genomes project

Breed	Frequency of the *RP1* frameshift allele in % (number of genomes per breed with available genotype)
Aberdeen Angus	1.8 (140)
Beef Booster Composite	2.1 (2)
Belgian Blue	5.0 (10)
Charolais	3.8 (39)
Gelbvieh	1.4 (36)
Holstein	1.8 (312)
Maine-Anjou	14.3 (7)
Normande	28.3 (23)
Red Angus	7.0 (28)
Run4	1.8 (1137)

The *RP1* frameshift allele was absent from the Brown Swiss (n = 59), Finnish red (n = 25), Hereford (n = 34), Jersey (n = 60), Limousine (n = 33), Montbéliarde (n = 28), Simmental (n = 215) and Swedish Red (n = 31) breeds, which each totalized more than 20 animals in run4 of the 1000 bull genomes project and from 12 additional breeds, which each totalized less than 20 animals

Table 2 Genotype frequencies for the *RP1* frameshift variant among 23 Normande founder bulls and 28 animals reported to ONAB for loss of vision

Genotype frequencies	Normande bulls in the 1000 bull genomes dataset (n = 23)	Loss of vision phenotype (n = 28)
Fs/Fs	8.7 % (n = 2)	50.0 % (n = 14)
Fs/Wt	43.5 % (n = 10)	35.7 % (n = 10)
Wt/Wt	47.8 % (n = 11)	14.3 % (n = 4)

Genotype frequencies calculated from the 23 Normande bulls available in run4 of the 1000 bull genomes project and from 28 cows declared to ONAB for loss of vision

ONAB French National Observatory of Bovine genetic Abnormalities, *Fs* frameshift allele, *Wt* wild type allele

The number of homozygous carriers (Fs/Fs) among affected animals is significantly larger than among the sequenced bulls (Chi^2-test p value = 0.00226)

Clinical and histological tests revealed symptoms of retinal degeneration in homozygous mutants

To gain better insight into the phenotypic consequences of this frameshift variant, we performed ocular tests on 23 pure and crossbred Normande cows of the same herd and for which genotype information was available. Genotypes were not disclosed to the veterinarian to exclude any bias of personal interpretation. All heterozygous and homozygous wild-type animals showed normal vision. Only a small proportion of them (three homozygous wild-type and two heterozygous) presented uni- or bilateral focal hyper-reflective areas in the tapetal fundus, which had no apparent consequences on their visual acuity. Among the four homozygous mutant animals, two heifers aged less than 3 years had normal vision and ocular tests. In contrast, two older animals aged 4.5 and 5.5 years presented respectively marked visual deficit and blindness, in spite of normal pupillary light reflexes. Their ocular fundi showed typical features of bilateral retinal degeneration with a heterogeneous color, multiple focal areas of hyper reflectivity in the tapetal area which could be coalescent, and a reduction in the caliber of retinal blood vessels (Fig. 3; Table 3). Thus, their phenotype was clearly distinct from the three homozygous wild-type individuals and the two heterozygous animals that displayed minor abnormalities of the ocular fundus.

Electroretinogram (ERG) performed on the oldest homozygous mutant confirmed the impairment of its retinal function with a lack of scotopic response and a reduced photopic response as compared with a wild-type control of the same age (Table 4).

Finally, to characterize this phenotype at the tissue level, we collected the retinas of two additional homozygous carriers (aged 8 years) and two control cows (one 6-year-old heterozygous Normande and one 8-year-old homozygous wild-type Holstein) after slaughter (Fig. 3). In concordance with previous analyses on the eyes' fundus and retinal function, histological analyses revealed a total absence of photoreceptor outer segments along with a marked thinning and disorganization of the outer nuclear layer with very few remaining nuclei.

Taken together, these results provide strong support that the *RP1* frameshift mutation causes a recessive loss of vision in bovine cattle. The phenotype observed is similar to the description in humans with a late onset of the disease due to progressive degeneration of the photoreceptors. Very few genetic conditions that affect eyesight have been reported in cattle [59] and, to our knowledge, this is the first time that a mutation causing retinal degeneration is reported in this species. Indeed, while, in the past, several cases of progressive retinal degeneration were reported in Holstein cows, their genetic etiology has not been confirmed so far [60, 61].

IBD analysis reveals a unique and ancestral mutation event

As previously mentioned, the frameshift mutation in *RP1* is not restricted to the Normande breed. So far, we have identified carriers in at least 12 cattle breeds: nine of the 15 breeds from the 1000 bull genomes dataset used in this study (Holstein, Charolais, Normande, Red Angus, Aberdeen-Angus, Gelbvieh, Beef Booster Composite, Maine-Anjou and Belgian Blue), and the Montbeliarde, Abondance and Vosgienne breeds based on EuroG10K genotyping results (Table 5). Since the mutation consists in the insertion of one adenosine in a polynucleotide stretch, which is more prone to mutation than other sites of the genome, we performed an IBD analysis to verify if only one ancestral mutation or multiple independent mutation events accounted for the wide distribution of this variant (see "Methods" section). In the 1000 bull genomes dataset, we identified a unique fragment of 88.6 kb (Chr14:23939194-24027957) that encompasses the mutation (Fig. 4) and was shared by all carriers (N = 38) but absent in non-carriers from the same breeds. This confirms the existence of a unique ancestral mutation event which, according to the size of the IBD segment, was dated back to approximately 565 generations, i.e. 2800 to 4000 years before present, considering that the generation interval can vary from 5 to 7 years (see "Methods" section).

The observation of this old variant at low to moderate frequencies in numerous bovine breeds could be explained by a combination of genetic drift and absence of or a very limited negative counter-selection due to the late onset of the defect. Nevertheless, the high frequency of this mutation observed in Normande cattle (27.7 % in the genotyped population for genomic selection) was particularly striking and led us to perform additional investigations to test for positive selection (either directly or mediated by hitch-hiking) in this breed.

First, we tested the association between the mutant allele and a series of 28 traits that are routinely evaluated (including production, morphology, reproduction

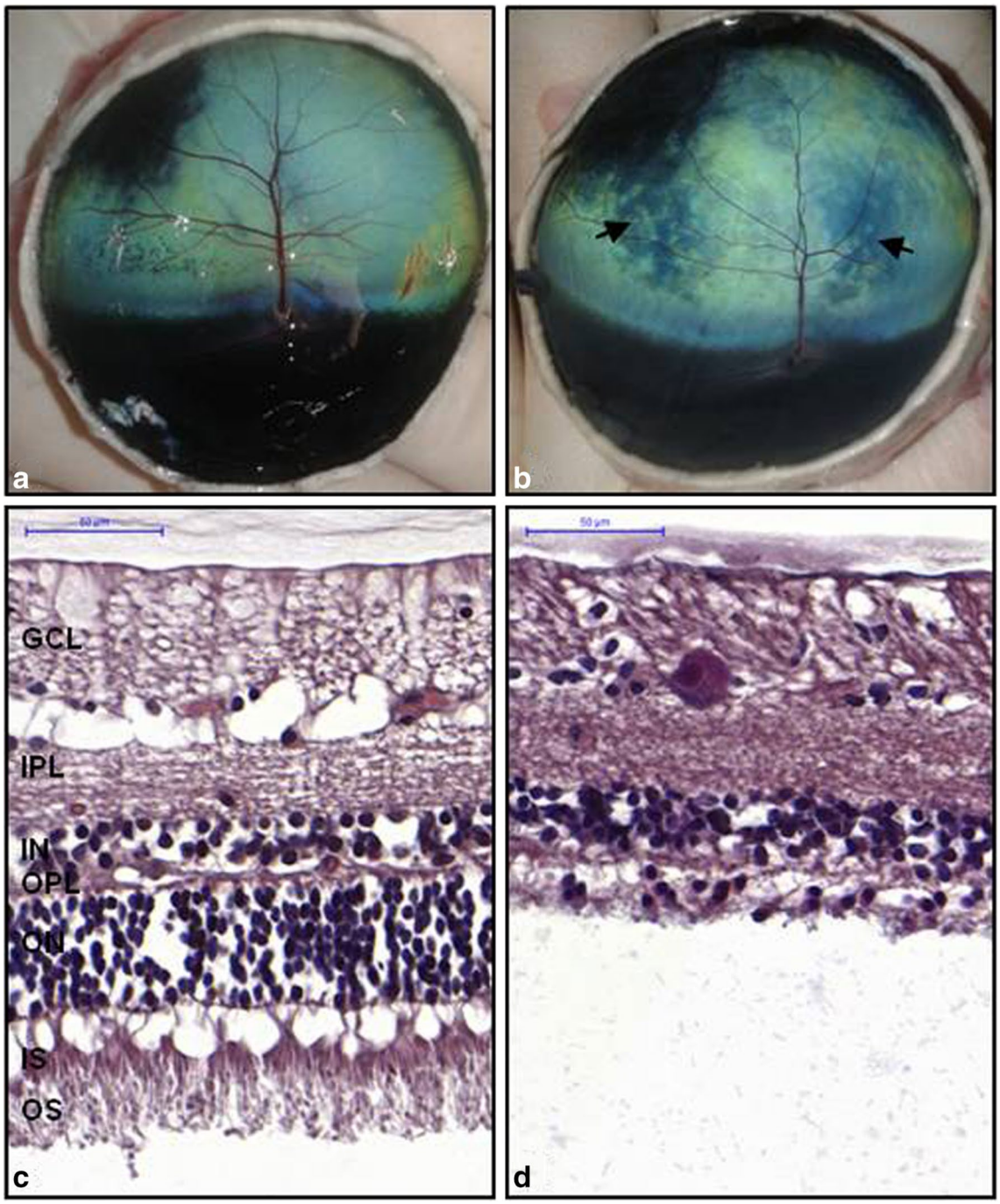

Fig. 3 Clinical and histological features of retinal degeneration in old Normande cows. **a** and **b** Eyes fundus from control *RP1* Fs/Wt (**a**) and affected *RP1* Fs/Fs (**b**) Normande cows. *Arrows* indicate hyperreflective areas and note the general reduction of the caliber of blood vessels in the affected animal. **c** and **d** Histological sections of the retina from the same control (**c**) and affected (**d**) animals. Note the total absence of inner and outer segments of photoreceptor cells and a marked thinning and disorganization of the outer nuclear layer confirming retinal degeneration in the Fs/Fs animal. (*GCL* ganglionic cells layer, *IPL* inner plexiform layer, *INL* inner nuclear layer, *OPL* outer plexiform layer, *ONL* outer nuclear layer, *IS* inner segment, *OS* outer segment. 5 µm sections of tissue embedded in paraffin and stained with HES. The choroid is not shown due to large artefactual gaps created by tissue preparation)

Table 3 Results of the ocular tests for the Normande dairy cattle examined

	Fs/Fs	Fs/Wt	Wt/Wt
Normal vision and eyes fundus	2 (<2 years)	7 (6.1 years)	7 (5.7 years)
Mild unilateral or bilateral focal retinal degeneration with preserved vision		2 (5.8 years)	3 (7.2 years)
Bilateral retinal degeneration with marked visual deficit	1 (4.5 years)		
Bilateral retinal degeneration with blindness	1 (5.5 years)		
Number of animals	4	9	10

Note that the pupillary light reflex was preserved in all the animals studied. The number of animals for each group and their average age in years (y) are presented

Fs frameshift allele, *Wt* wild type allele

Table 4 Electroretinogram results i.e. values of the amplitudes and the culminating times of a- and b-waves in one control and one affected animal

	Amplitude (μV)		Culminating time (msec)	
	a-wave	b-wave	a-wave	b-wave
Rod response				
Control	–	429	–	52
Affected	–	–	–	–
Mixed response				
Control	80.9	274	17	60
Affected	–	–	–	–
Cone response				
Control	36.7	275	13	24
Affected	15.6	43.6	17	27

Table 5 Genotype frequencies for the *RP1* frameshift polymorphism from the EuroG10K genotyping results

Breed	Wt/Wt	Fs/Wt	Fs/Fs	MAF (%)
Abondance	1633	1	0	0.03
Brown Swiss	296	0	0	0.00
Tarentaise	698	0	0	0.00
Simmental	579	0	0	0.00
Montbéliarde	40,188	359	1	0.45
Normande	6294	4915	897	27.71
Vosgienne	505	2	0	0.20
Holstein	51,640	1627	12	1.55
Red pied lowland	1005	0	0	0.00

Fs frameshift allele, *Wt* wild type allele

and health) as well as three coat color phenotypes. A strong association was found only with two udder traits, i.e. front teat distance and teat orientation, with an unfavorable effect of the mutant allele. Some putative effects of lower magnitude were also observed on fat and protein contents (Table 6). None of these effects can explain the high frequency of the mutant allele.

Second, using Illumina EuroG10K SNP genotyping data or phased Illumina BovineSNP50 haplotypes (see "Methods" section), we estimated the allelic frequencies over the last 40 years within the AI bull population (Fig. 5). Interestingly, the frequency of the *RP1* frameshift mutation showed a progressive decrease (from 40 to 27 %) during this period. Thus, the increase in frequency of the mutant allele in the Normande breed is more ancient and most probably results from a founder effect that was favored by the advent of AI in the 1950s. Because of the late onset of the defect and because the dams of the future AI bulls are primarily selected among young cows to reduce generation intervals and increase the annual genetic gain, it is unlikely that the decrease in allelic frequency is caused by selection against blindness. A possible explanation of this negative trend is the association of the mutant allele with udder morphology and the strong selection on this trait in the last 50 years. Indeed, the original udder morphology of Normande cows was not adapted to machine milking and was gradually improved over time through drastic selection. While this *RP1* mutation has very limited economic impact, it has major implications in terms of animal welfare and human safety. Indeed, with a frequency of 27 % in 2015, about one in every 14 Normande animals will become progressively blind and be subject to increased stress and fear, as we observed during sampling. This also means that each farmer possesses more than one homozygous carrier and has an increased risk of being injured by a startled animal. The identification of this mutation and its incorporation into the EuroG10K SNP chip used for genomic selection provide the basis for its active counter-selection.

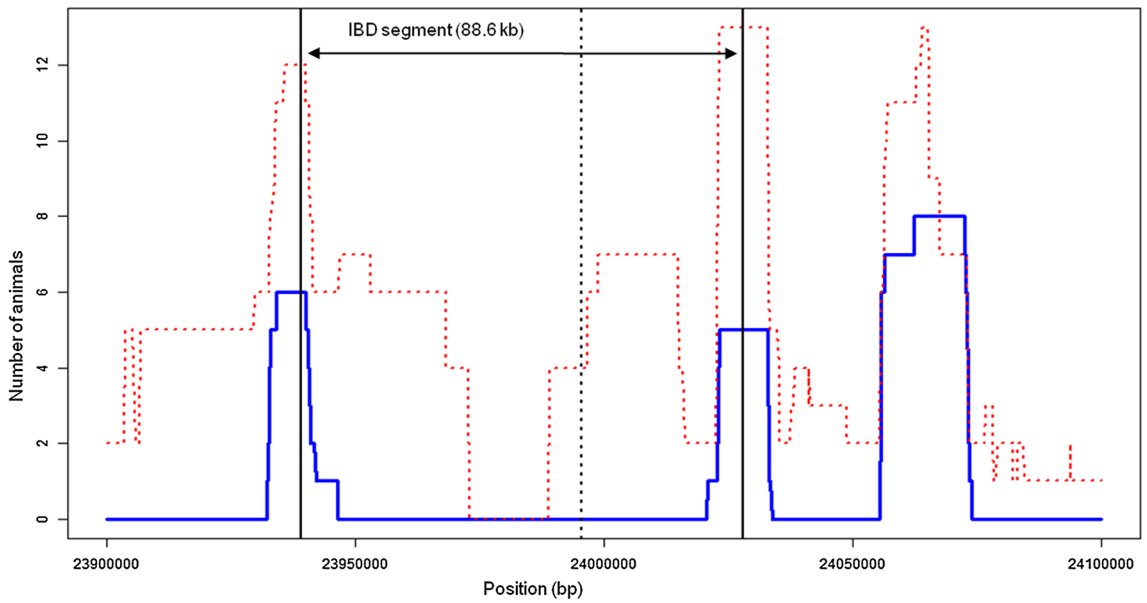

Fig. 4 IBD analysis around the *RP1* frameshift mutation. *Blue line* number of animals carrying the *RP1* frameshift allele but showing at least 5 % of homozygous genotypes for the alternative allele as compared with the *RP1* Fs/Fs Normande bull Diametre in sliding windows of 100 SNPs. Thirty-five Fs/Wt and three Fs/Fs animals were considered. Successive windows around the mutation for which this count is null defines an 88.6-kb IBD segment common to all carriers. The same calculation was applied to a control group of 38 randomly chosen non-carrier bulls from the same breeds (*red dashed line*). Note that the segment around the mutation is not conserved in this group. The *black dashed* line indicates the localization of the mutation

Table 6 Results of association studies between the *RP1* frameshift mutation and 31 traits routinely evaluated for 11,986 Normande cows

Traits	Effect	Standard error	p value (Bonferroni)
Front teat distance	−0.106	0.021	1.2E−05
Teat orientation	−0.076	0.019	2.4E−03
Milk protein content	−0.069	0.019	9.5E−03
Milk fat content	−0.132	0.037	9.9E−03

Only significant results after Bonferroni correction are presented. Details on the 31 different traits studied are presented in "Methods" section. Front teat distance and teat orientation are scored from 1 to 9. Milk protein and fat content are expressed in g/L

Conclusions

In this work, we have drawn a list of putative deleterious mutations which are not rare (frequency higher than 5 %) in at least one of 15 investigated European bovine breeds. We showed that these variants, which likely represent a substantial part of the genetic load of domestication in cattle, did not accumulate randomly. Our results reveal that mutations in genes, which are involved in cognition or sensorial functions for which little or no selection pressure exists in domesticated animals, were more tolerated than mutations that affect other systems, which are of primary importance for production, reproduction and survival. Among these variants, we described an ancestral frameshift mutation in *RP1* which segregates in numerous breeds and causes progressive retinal degeneration. To our knowledge, this is the most ancient and widespread mutation causing a recessive genetic defect in cattle reported to date. This example illustrates that our approach can help to unravel variants that are yet to be discovered and are the cause of unselected but debilitating phenotypes in domestic animals. We are confident that the phenotypic characterization of a number of the variants reported here will offer interesting results in the near future.

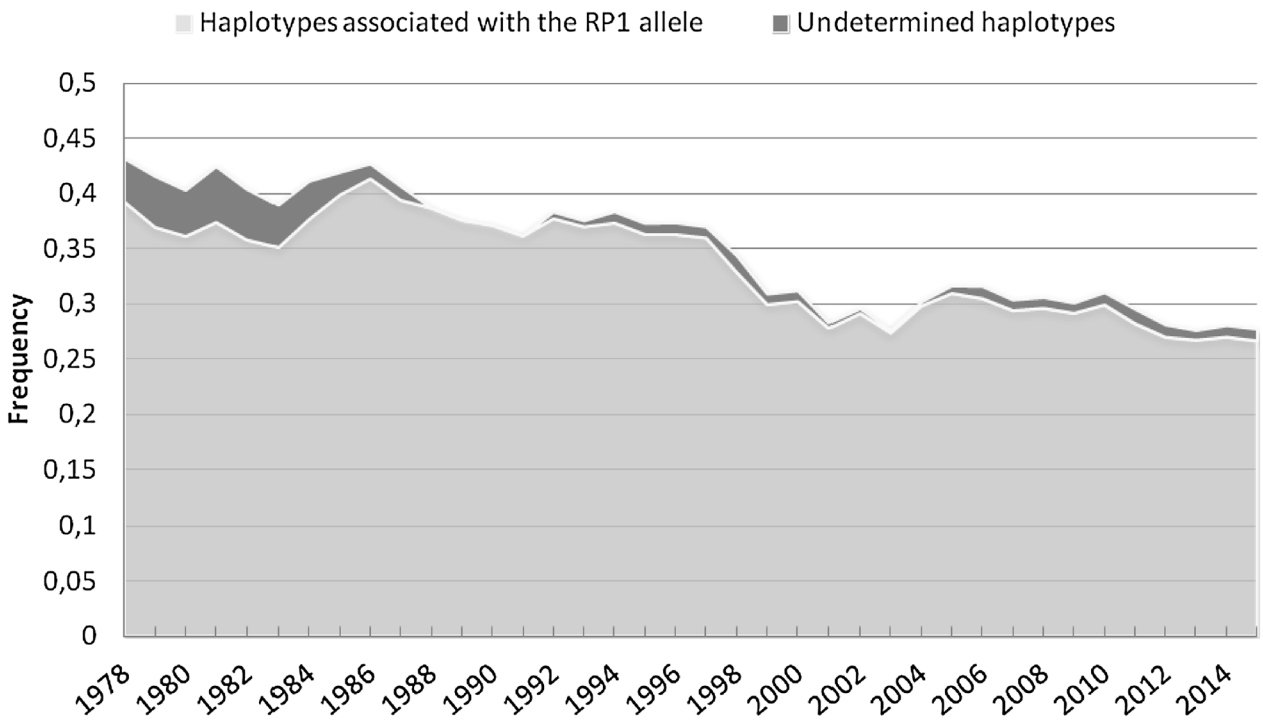

Fig. 5 Changes in the frequency of the *RP1* frameshift allele within the AI Normande bulls population. Frequencies were calculated using a haplotype test approach for sliding windows of 7 years for AI bulls born from 1970 to 2015. Undetermined haplotypes correspond to haplotypes which were observed in both Fs/Fs and Wt/Wt animals genotyped for the *RP1* frameshift mutation with the Illumina EuroG10K SNP chip (see "Methods" section)

Additional files

Additional file 1: Table S1. Details on animals used in the phenotypic characterization of the *RP1* frameshift. This table provides details on the animals used for each analysis in this study. AAN: Aberdeen-Angus, BBB: Belgian Blue, BBR: Beef Booster Composite, CHA: Charolais, GVH: Gelbvieh, HOL: Holstein, NOR: Normande, NOR*HOL: Normande*Holstein crossbred RAN: Red Angus, RDP: Maine-Anjou.

Additional file 2: Table S2. Details on non-rare putative deleterious variants selected in 15 breeds from the 1000 bull genomes run4 dataset. This file contains the list of the 2489 variants selected in this study. For each of them, we indicate: (i) the frequency in each breed, (ii) the functional consequence on the protein, and (iii) the genetic syndromes associated with mutations within the same gene in human (Online Mendelian Inheritance in Man, OMIM; http://www.omim.org) and mouse (Mammalian Phenotypes; http://www.informatics.jax.org). AAN: Aberdeen-Angus, BBR: Beef Booster Composite, BSW: Brown Swiss, CHA: Charolais, GVH: Gelbvieh, HER: Hereford, HOL: Holstein, JER: Jersey, LIM: Limousine, MON: Montbéliarde, NOR: Normande, RAN: Red Angus, RDCF: Finnish Red, RDCS: Swedish Red, and SIM: Simmental.

Additional file 3: Tables S3, S4, S5, S6 and S7. Additional information on the results of the gene enrichment analysis performed with Ingenuity Pathway Analysis. This file contains five tables providing additional details on the results of the gene enrichment analysis: a summary of the significant diseases and disorders annotations (Table S3), a summary of significant physiological system development and function annotations (Table S4), a summary of significant molecular and cellular functions annotations (Table S5), the key-word attribution to each significant functional annotation conserved for the analysis (Table S6) and the significant canonical pathways (Table S7).

Authors' contributions

AC conceived and coordinated the study. AC, PM, SC and DB designed the study. SC, DD and ED performed clinical examinations. PM and AB performed histological analyses. PM, AC and MB analyzed WGS data. PM and AC performed gene enrichment analysis. PM, AC, SB and MC participated to sampling. MC, SB and CDB provided phenotype or pedigree information. CG, MCD and AAB performed DNA extraction and genotyping by PCR-sequencing. LG performed quality control of the SNP chip genotyping data and provided DNA samples from carriers of the *RP1* mutation. PM, A and SF performed phasing and haplotype testing. AM performed imputation and association studies. PM, AC, SC and DB wrote the manuscript. All authors read and approved the final manuscript.

Author details

[1] UMR 1313 GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France. [2] ALLICE, 149 rue de Bercy, 75595 Paris Cedex 12, France. [3] Ecole Nationale Vétérinaire d'Alfort, Unité d'Ophtalmologie, Université Paris-Est, 7 avenue du Général de Gaulle, 94704 Maisons-Alfort Cedex, France. [4] Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark. [5] UMR 1198, Biologie du Développement et Reproduction, INRA, 78350 Jouy-en-Josas, France. [6] Origenplus, 38 rue de la Mérillière, 61300 L'Aigle, France. [7] UE 326, Domaine Expérimental du Pin-au-haras, INRA, 61310 Exmes, France. [8] Labogena DNA, Domaine de Vilvert, 78350 Jouy-en-Josas, France. [9] IDELE, 149 rue de Bercy, 75595 Paris Cedex 12, France.

Acknowledgements

The authors are grateful to the breeders of the "GAEC de l'Araucaria" for their hospitality, Maelle Philippe (EVOLUTION) for providing samples, Albéric Valais (Normande breed Organization) for providing phenotype information, and the partners of the 1000 bull genomes consortium for the excellent collaboration. This study is part of the BOVANO project (ANR-14-CE19-0011) funded by the French Agence Nationale de la Recherche and Apisgene. P. Michot is recipient of a PhD. Grant from ALLICE and Apis Gène.

Competing interests
The authors declare that they have no competing interests.

Availability of data and material
The datasets supporting the conclusions of this article are included within the article and its additional files.

Funding
P. Michot is recipient of a PhD. Grant from ALLICE and Apis Gène. This study is part of the BOVANO project (ANR-14-CE19-0011) funded by the French Agence Nationale de la Recherche and Apisgene.

Received: 2 May 2016 Accepted: 26 July 2016
Published online: 10 August 2016

References

1. Lu J, Tang T, Tang H, Huang J, Shi S, Wu C. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. 2006;22:126–31.
2. Cruz F, Vila C, Webster MT. The legacy of domestication: accumulation of deleterious mutations in the dog genome. Mol Biol Evol. 2008;25:2331–6.
3. Nabholz B, Sarah G, Sabot F, Ruiz M, Adam H, Nidelet S, et al. Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*Oryza glaberrima*). Mol Ecol. 2014;23:2210–27.
4. Koenig D, Jimenez-Gomez JM, Kimura S, Fulop D, Chitwood DH, Headland LR, et al. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proc Natl Acad Sci USA. 2013;110:2655–62.
5. Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc Natl Acad Sci USA. 2014;111:E5661–9.
6. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, et al. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet. 2008;40:449–54.
7. Michot P, Fantini O, Braque R, Allais-Bonnet A, Saintilan R, Grohs C, et al. Whole-genome sequencing identifies a homozygous deletion encompassing exons 17 to 23 of the integrin beta 4 gene in a Charolais calf with junctional epidermolysis bullosa. Genet Sel Evol. 2015;47:37.
8. Grobet L, Martin LJ, Poncelet D, Pirottin D, Brouwers B, Riquet J, et al. A deletion in the bovine *myostatin* gene causes the double–muscled phenotype in cattle. Nat Genet. 1997;17:71–4.
9. Arthur PF, Makarechian M, Price MA. Incidence of dystocia and perinatal calf mortality resulting from reciprocal crossing of double-muscled and normal cattle. Can Vet J. 1988;29:163–7.
10. Fasquelle C, Sartelet A, Li W, Dive M, Tamma N, Michaux C, et al. Balancing selection of a frame-shift mutation in the *MRC2* gene accounts for the outbreak of the crooked tail syndrome in Belgian Blue cattle. PLoS Genet. 2009;5:e1000666.
11. Sartelet A, Druet T, Michaux C, Fasquelle C, Géron S, Tamma N, et al. A splice site variant in the bovine *RNF11* gene compromises growth and regulation of the inflammatory response. PLoS Genet. 2012;8:e1002581.
12. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
15. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26:2069–70.
16. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
17. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.
18. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, et al. A network-based analysis of systemic inflammation in humans. Nature. 2005;437:1032–7.
19. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000;132:365–86.
20. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, et al. novoSNP, a novel computational tool for sequence variation discovery. Genome Res. 2005;15:436–42.
21. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.
22. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
23. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84:210–23.
24. Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, et al. Genomic selection in French dairy cattle. Anim Prod Sci. 2016;52:115–20.
25. Boitard S, Boussaha M, Capitan A, Rocha D, Servin B. Uncovering adaptation from sequence data: lessons from genome resequencing of four cattle breeds. Genetics. 2016;203:433–50.
26. Grobet L, Poncelet D, Royo LJ, Brouwers B, Pirottin D, Michaux C, et al. Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. Mamm Genome. 1998;9:210–3.
27. Fritz S, Capitan A, Djari A, Rodriguez SC, Barbat A, Baur A, et al. Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in *GART*, *SHBG* and *SLC37A2*. PLoS One. 2013;8:e65550.
28. Sonstegard TS, Cole JB, VanRaden MP, Van Tassell CP, Null DJ, Schroeder SG, et al. Identification of a nonsense mutation in *CWC15* associated with decreased reproductive efficiency in Jersey cattle. PLoS One. 2013;8:e54872.
29. Lunden A, Marklund S, Gustafsson V, Aderdersson L. A nonsense mutation in the *FMO3* gene underlies fishy off-flavor in cow's milk. Genome Res. 2002;12:1885–8.
30. Berry SD, Davis SR, Beattie EM, Thomas NL, Burrett AK, Ward HE, et al. Mutation in bovine-*carotene oxygenase 2* affects milk color. Genetics. 2009;182:923–6.
31. Robertson NG, Skvorak AB, Yin Y, Weremowicz S, Johnson KR, Kovatch KA, et al. Mapping and characterization of a novel cochlear gene in human and in mouse: a positional candidate gene for a deafness disorder, DFNA9. Genomics. 1997;46:345–54.
32. Cohen-Salmon M, Mattei MG, Petit C. Mapping of the *otogelin* gene (*OTGN*) to mouse chromosome 7 and human chromosome 11p14.3: a candidate for human autosomal recessive nonsyndromic deafness DFNB18. Mamm Genome. 1999;10:520–2.
33. Yariz KO, Duman D, Seco CZ, Dallman J, Huang M, Peters TA, et al. Mutations in *OTOGL*, encoding the inner ear protein otogelin-like, cause moderate sensorineural hearing loss. Am J Hum Genet. 2012;91:872–82.
34. Wang A, Liang Y, Fridell RA, Probst FJ, Wilcox ER, Touchman JW, et al. Association of unconventional myosin *MYO15* mutations with human nonsyndromic deafness DFNB3. Science. 1998;280:1447–51.
35. Verpy E, Masmoudi S, Zwaenepoel I, Leibovici M, Hutchin TP, Del Castillo I, et al. Mutations in a new gene encoding a protein of the hair bundle cause non-syndromic deafness at the *DFNB16* locus. Nat Genet. 2001;29:345–9.
36. Motazacker MM, Rost BR, Hucho T, Garshasbi M, Kahrizi K, Ullmann R, et al. A defect in the *ionotropic glutamate receptor 6* gene (*GRIK2*) is associated with autosomal recessive mental retardation. Am J Hum Genet. 2007;81:792–8.
37. Kandaswamy R, McQuillin A, Curtis D, Gurling H. Allelic association, DNA resequencing and copy number variation at the metabotropic glutamate receptor *GRM7* gene locus in bipolar disorder. Am J Med Genet B Neuropsychiatr Genet. 2014;165:365–72.
38. Shifman MI, Selzer ME. Differential expression of class 3 and 4 semaphorins and netrin in the lamprey spinal cord during regeneration. J Comp Neurol. 2007;501:631–46.

39. Kantor DB, Chivatakarn O, Peer KL, Oster SF, Inatani M, Hansen MJ, et al. Semaphorin 5A is a bifunctional axon guidance cue regulated by heparan and chondroitin sulfate proteoglycans. Neuron. 2004;44:961–75.
40. Mosca-Boidron AL, Gueneau L, Huguet G, Goldenberg A, Henry C, Gigot N, et al. A de novo microdeletion of *SEMA5A* in a boy with autism spectrum disorder and intellectual disability. Eur J Hum Genet. 2015;24:838–43.
41. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet. 2008;40:1056–8.
42. Breitenkamp AF, Matthes J, Nass RD, Sinzig J, Lehmkuhl G, Nürnberg P, et al. Rare mutations of *CACNB2* found in autism spectrum disease-affected families alter calcium channel function. PLoS One. 2014;9:e95579.
43. Stefansson H, Petursson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, et al. Neuregulin 1 and susceptibility to schizophrenia. Am J Hum Genet. 2002;71:877–92.
44. Brown SM, Clapcote SJ, Millar JK, Torrance HS, Anderson SM, Walker R, et al. Synaptic modulators Nrxn1 and Nrxn3 are disregulated in a Disc1 mouse model of schizophrenia. Mol Psychiatry. 2011;16:585–7.
45. Jeyabalan N, Clement JP. SYNGAP1: mind the Gap. Front Cell Neurosci. 2016;10:32.
46. Carneiro M, Rubin CJ, Di Palma F, Albert FW, Alfoldi J, Barrio AM, et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science. 2014;345:1074–9.
47. Rubin CJ, Megens HJ, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci USA. 2012;109:19529–36.
48. Rubin CJ, Zody MC, Eriksson JR, Meadows J, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464:587–91.
49. Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, et al. Classic selective sweeps revealed by massive sequencing in cattle. PLoS Genet. 2014;10:e1004148.
50. Ramey HR, Decker JE, McKay SD, Rolf MM, Schnabel RD, Taylor JF. Detection of selective sweeps in cattle using genome-wide SNP data. BMC Genomics. 2013;14:382.
51. Flori L, Fritz S, Jaffrezic F, Boussaha M, Gut I, Heath S, et al. The genome response to artificial selection: a case study in dairy cattle. PLoS One. 2009;4:e6595.
52. Mayilyan KR. Complement genetics, deficiencies, and disease associations. Protein Cell. 2012;3:487–96.
53. Diodato D, Ghezzi D, Tiranti V. The mitochondrial aminoacyl tRNA synthetases: genes and syndromes. Int J Cell Biol. 2014;2014:787956.
54. Pierce EA, Quinn T, Meehan T, McGee TL, Berson EL, Dryja TP. Mutations in a gene encoding a new oxygen-regulated photoreceptor protein cause dominant retinitis pigmentosa. Nat Genet. 1999;22:248–54.
55. Audo I, Mohand-Saïd S, Dhaenens CM, Germain A, Orhan E, Antonio A, et al. RP1 and autosomal dominant rod-cone dystrophy: novel mutations, a review of published variants, and genotype–phenotype correlation. Hum Mutat. 2011;33:73–80.
56. Liu Q, Collin RW, Cremers FP, den Hollander AI, van den Born LI, Pierce EA. Expression of wild-type Rp1 protein in Rp1 knock-in mice rescues the retinal degeneration phenotype. PLoS One. 2012;7:e43251.
57. El Shamieh S, Boulanger-Scemama E, Lancelot ME, Antonio A, Démontant V, Condroyer C, et al. Targeted next generation sequencing identifies novel mutations in RP1 as a relatively common cause of autosomal recessive rod-cone dystrophy. Biomed Res Int. 2015;2015:485624.
58. Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. Lancet. 2006;368:1795–809.
59. Murgiano L, Jagannathan V, Calderoni V, Joechler M, Gentile A, Drögemüller C. Looking the cow in the eye: deletion in the *NID1* gene is associated with recessive inherited cataract in Romagnola cattle. PLoS One. 2014;9:e110628.
60. Bradley R, Terlecki S, Clegg FG. The pathology of a retinal degeneration in Friesian cows. J Comp Pathol. 1982;92:69–83.
61. Stehmann SM, Rebhun WC, Riis RC. Progessive retinal atrophy in related cattle. The Bovine Practitioner. 1987;20:195–7.

CHAPITRE 7. DISCUSSION GENERALE ET PERSPECTIVES

Au cours de ma thèse, j'ai appliqué différentes stratégies d'exploitation des données de génotypage et de séquences dans le but d'identifier des mutations récessives responsables d'anomalies génétiques chez le bovin.

Nous avons abordé des approches usuelles de cartographie génétique déjà largement utilisées, et des approches nouvelles de génétique inverse dont le développement est possible chez le bovin grâce à la quantité de séquences et de données de typages recueillies sur les différentes populations dans le cadre de la sélection génomique et des travaux de recherche. En cela, mes travaux reflètent les évolutions récentes des méthodologies et l'intégration de l'utilisation systématique des données de séquence de génome. Ces données à haut débit offrent un volume considérable de variants identifiés dont on peut tirer parti dans la découverte de nouvelles mutations potentiellement responsables de phénotypes délétères chez les animaux d'élevage.

Les résultats de chaque étude ont été discutés au sein des chapitres. Ici, je présenterai un bilan général des résultats et discuterai des avantages et limites des approches utilisées pour l'exploitation des données de séquence. Je développerai ensuite des perspectives à court terme pour faire suite aux travaux entrepris dans le cadre de cette thèse, et à plus long terme pour une utilisation plus efficace des données de séquence de génomes complets dans les années futures. Enfin, je présenterai l'intérêt d'étudier les anomalies génétiques pour les filières d'élevage, mais aussi dans un contexte d'amélioration des connaissances du génome et de la fonction des gènes.

I. Bilan et limites de l'utilisation des données de séquence pour l'identification des mutations responsables d'anomalies génétiques

A. Détection des mutations responsables d'anomalies émergentes

En utilisant des approches usuelles, allant du phénotype à l'identification d'une mutation candidate, j'ai pu étudier deux anomalies génétiques récessives dont l'émergence a été détectée par l'ONAB.

Grâce à une cartographie par homozygotie sur marqueurs de la puce BovineSNP50 d'Illumina et à l'analyse des données de séquence de deux animaux atteints d'épilepsie idiopathique en race Parthenaise, nous avons identifié une région de 1,2 Mb sur le chromosome 24 puis une mutation d'épissage dans le gène *MTCL1*, seule mutation potentiellement délétère dans l'intervalle associé au syndrome. A partir des données de séquence de génome complet d'un seul veau Charolais atteint

d'épidermolyse bulleuse jonctionnelle, nous avons également identifié une délétion affectant les exons 17 à 23 du gène *ITGB4* en recherchant des mutations délétères dans des gènes candidats situés dans des régions d'homozygotie.

Au travers de ces deux études, nous montrons que l'accès aux données de séquence de génome complet et l'annotation bioinformatique de ces génomes permet d'identifier rapidement les mutations candidates. Le séquençage de génomes nous a permis de détecter l'ensemble des variants portés par chaque individu et, en fonction des cas, de se détacher :

- de la cartographie génétique préalable du locus associé, comme nous le montrons avec les travaux sur l'EBJ
- de l'approche par gène candidat positionnel, comme dans notre étude de l'épilepsie où aucun gène candidat évident n'était présent dans la région.

Dans chaque cas, nous disposions d'un pedigree suffisamment détaillé pour émettre l'hypothèse d'un déterminisme génétique récessif, en particulier pour l'EBJ avec une consanguinité forte chez les animaux atteints (issus d'accouplements consanguins père-fille et demi-frères-sœurs). Ceci nous a orientés vers une stratégie d'identification des régions homozygotes chez les animaux atteints et l'application de filtres bioinformatiques permettant la détection des variations candidates en accord avec ces hypothèses.

Deux points de discussion peuvent être abordés dans le choix du séquençage des génomes complets des individus et des filtres appliqués pour l'identification des mutations : l'intérêt d'un WGS et la stratégie de filtre contre les données WGS de plusieurs centaines d'individus.

1. Pourquoi avoir choisi le séquençage complet du génome des individus atteints ?

Dans l'étude des syndromes d'épilepsie et de l'EBJ, nous avons fait le choix de réaliser le séquençage de l'intégralité des génomes d'individus atteints pour rechercher les mutations candidates. Pourquoi avoir privilégié ce choix ?

En effet, la majorité des anomalies génétiques à transmission mendélienne (ie. mono-géniques) sont liées à une modification de la partie codante d'un gène qui impacte ensuite la protéine produite. Dans le cas de l'épilepsie, nous avons localisé la mutation causale dans un intervalle de quelques mégabases sur le chromosome 24. Deux approches auraient été plus économes en séquençage. Nous aurions pu cibler les séquences codantes et donc séquencer uniquement l'exome ; alternativement, nous aurions pu cibler uniquement la région d'intérêt et réaliser la capture et le séquençage de cette petite région au lieu de considérer le génome entier. Le séquençage de l'exome est une stratégie très utilisée chez l'homme pour la recherche des mutations associées à une anomalie génétique (Bamshad *et al.,* 2011). Cette stratégie a été appliquée efficacement chez le bovin, par exemple pour l'identification de la mutation HH3 (McClure *et al.,* 2014) ainsi que dans la récente étude à grande échelle de Charlier *et al* (2016).

La capture et le séquençage d'une région d'intérêt ont aussi été privilégiés au début de l'utilisation des NGS. Cette stratégie a servi par exemple pour l'identification de la mutation dans le gène SUOX (*Sulfite Oxidase)* responsable de l'arachnomelia en race Brune (Drögemüller *et al.*, 2010) ou encore l'identification de la mutation dans le gène *CEP250* (Centrosomal Protein 250) associée au syndrome SHCG en Montbéliarde (Floriot *et al.*, 2015). L'avantage de ces deux techniques est de pouvoir diminuer le coût lié au séquençage tout en augmentant sur la région ciblée la qualité de la séquence obtenue en augmentant la profondeur de séquençage. Ces techniques permettent également d'envisager de séquencer plus d'individus et donc d'augmenter la puissance de détection de la mutation causale. Cependant, le coût d'une séquence de génome complet a beaucoup diminué tout en fournissant une couverture en général suffisante pour capter les variants. L'intérêt économique de la capture spécifique de la région d'intérêt, très fort initialement, s'est beaucoup réduit aujourd'hui, et l'approche génome complet a l'avantage d'être générique et donc plus rapide à mettre en œuvre.

La capture d'exome est devenue générique chez l'homme et elle est bien maitrisée dans certaines équipes. Elle aurait pu constituer une alternative pour nous. Toutefois, nous n'avions pas accès aux outils techniques et notre laboratoire n'a pas choisi cette stratégie. Les raisons essentielles en sont les suivantes :

(a) même si les variants dans le codant sont les plus probables, on ne peut pas exclure les autres régions du génome et la séquence du génome complet permet d'aborder cet aspect si nécessaire dans une seconde approche ;

(b) la séquence du génome complet permet de façon beaucoup plus fiable de détecter les variants structuraux, comme dans le cas de l'EBJ en race Charolaise ;

(c) enfin, une stratégie de séquençage du génome complet permet de capitaliser les séquences entre projets et de constituer une base de données générale décrivant le maximum de variants. Ainsi, notons par exemple qu'un individu atteint de l'anomalie A peut être un témoin pour l'anomalie B localisée à un autre endroit du génome (voir § 2) ; il contribue aussi à la description du déséquilibre de liaison sur le génome et donc à la qualité d'imputation. Ceci est d'autant plus vrai que les cas appartiennent à des populations à faible effectif d'animaux séquencés, comme par exemple pour la race Parthenaise.

2. Filtrer des variations contre les données de séquence disponibles dans d'autres races ?

Les filtres bioinformatiques appliqués pour isoler les variations candidates comparent les variants identifiés chez un animal atteint à ceux identifiées chez un individu sain. L'objectif de cette étape est de réduire au maximum le nombre de mutations associées au phénotype. Dans nos approches, nous avons utilisé comme contrôle les données de séquence de génome complet de taureaux d'IA, selon une stratégie identique à celle appliquée par Daetwyler *et al* (2014). Dans le cas des anomalies récessives,

d'autres individus peuvent être séquencés pour servir de contrôle comme par exemple des frères/sœurs non-atteints.

La meilleure stratégie pour identifier la mutation candidate serait de disposer de la séquence d'un animal porteur de l'haplotype ancestral, c'est-à-dire l'haplotype identique à celui porté par les animaux atteints mais sans la mutation délétère. En utilisant la structure des populations bovines, les informations de génotypes et de pedigree, il est parfois possible d'identifier ce type d'individus au sein de la population ou du pedigree de l'anomalie (Drögemüller *et al.*, 2010). En théorie, en comparant les variants ségrégant avec l'haplotype ancestral et ceux ségrégant avec l'haplotype porteur de la mutation, on ne devrait identifier qu'une seule différence : la mutation candidate. Cette situation très favorable n'est rencontrée que pour des mutations très récentes et un haplotype relativement fréquent dans la population. Un cas particulièrement favorable est celui d'une neo mutation dominante, donc absente du génome des parents. En cas de mosaïcisme du parent, on peut également comparer deux produits portant le même haplotype, l'un atteint et l'autre sain ; cet haplotype ne différant essentiellement que par la neo mutation. Dans le cas d'une anomalie récessive, les porteurs du même haplotype sain sont beaucoup plus éloignés et un screening par génotypage est nécessaire pour les détecter.

Une alternative est d'utiliser un grand nombre d'individus contrôles, décrivant la plus grande partie des polymorphismes existant à l'échelle de chaque race. Aujourd'hui, avec les travaux de séquençage de génomes, on dispose d'une très grande base de variations identifiées sur des animaux de différentes races. En 2015, dans son Run 5, le consortium « 1000 bull genomes » avait produit 1577 séquences de génomes complets de 33 races bovines taurines et identifié 39,7 millions de variants. Par exemple, pour l'épilepsie les variants identifiés chez les deux animaux atteints ont été filtrés contre les variants identifiés chez plus de 1323 individus non atteints. Ainsi, nous avons pu réduire le nombre de variants dans l'intervalle de 1269 à seulement neuf candidats positionnels, dont la seule mutation avec un effet délétère fort située dans le gène *MTCL1*. On atteint donc des niveaux d'efficacité d'identification des candidats très élevés.

B. Approches de génétique inverse

Au cours de cette thèse, nous avons tenté de mettre en place diverses méthodes, fondées sur l'exploitation des variants identifiés dans les données de séquençage pour caractériser des mutations causales pour lesquelles nous ne disposions pas ou peu d'informations phénotypiques.

1. Intérêt des approches inverses : se détacher de la déclaration des phénotypes

En continuité des travaux de Fritz *et al.*, (2013), nous proposons une nouvelle mutation candidate dans le gène *PFAS* associée à l'haplotype de mortalité embryonnaire MH1. Son identification résulte de la

combinaison des approches d'études haplotypiques et de screening des variants identifiés dans les données de séquence des individus porteurs.

Par ailleurs, nous avons exploré des approches de génétique inverse pour caractériser des variations pour lesquelles les outils bioinformatiques prédisent un effet potentiellement délétère sur la fonction des gènes dans lesquels elles sont situées. En appliquant deux hypothèses différentes, l'une fondée sur les variants rares race spécifiques et l'autre sur les variants fréquents, nous avons mis en évidence deux nouvelles mutations en race Normande. La première située dans le gène *CAD* est très probablement associée à une mortalité embryonnaire, et la seconde dans le gène *RP1* associée à une perte progressive de la vision chez les animaux adultes.

Nos travaux de caractérisation des variants fréquents et la confirmation de l'effet de la mutation dans *RP1* suggèrent fortement que des variants délétères fréquents ségrégent chez le bovin et touchent des voies de gènes dont les fonctions sont moins soumises à la pression de sélection exercée en condition d'élevage comme la vision, le comportement et la cognition (*voir chapitre 6 ; Michot et al., 2016*). Ce sont des cibles peu étudiées jusqu'à présent car leur impact semble modéré et, de ce fait, difficilement identifiables par le réseau de surveillance actuel.

Le développement des approches inverses, à partir des données de génotype ou de séquence, permettent de se détacher des déclarations de phénotypes anormaux aux observatoires nationaux. De nombreuses études ont montré leur intérêt pour identifier des mutations responsables de morts embryonnaires, mais aussi repérer des variants non létaux, qui ont parfois un faible impact sur la carrière de l'individu. On peut citer comme exemple la mutation du gène *RP1* conduisant à une perte progressive de vision en race Normande ou une délétion dans le gène codant pour la mélanophilin (MLPH) qui affecte la coloration des individus en race Blanc Bleu Belge (Li *et al.*, 2016).

Comme suggéré dans les études déjà mises en place chez l'homme, l'utilisation des données de séquence de génomes ou d'exomes permet de renverser la situation en ayant accès à un grand nombre de variants dont on cherche à expliquer l'effet sur le phénotype (Sulem *et al.*, 2015). En revanche, même si l'on se détache de la déclaration d'un phénotype anormal comme point de départ, la description clinique fine des animaux homozygotes pour la mutation reste une phase essentielle. Comme nous l'avons démontré avec l'étude du variant du gène RP1, la mise en place d'une description clinique est facilitée par la possibilité de cibler les animaux d'intérêt dans les élevages et d'organiser les examens à appliquer en fonction du gène d'intérêt et du phénotype attendu (Michot *et al.*, 2016).

2. Approche de génétique inverse et caractérisation de variants délétères race-spécifiques: les limites de l'approche appliquée

Dans le chapitre 5, la stratégie d'étude des variations que nous avons appliquée est comparable à celle employée par Charlier *et al.,* (2016). A partir du traitement des données de séquence, nous avons identifié des SNP et petits InDels affectant les régions codantes du génome, prédits comme délétères

pour les protéines correspondantes, jamais observés à l'état homozygote et spécifiques d'une race donnée. Dans un second temps, nous avons testé ces variations sur la population dans une étape d'identification de gènes candidats par génotypage. Il est difficile de chiffrer l'efficacité de notre méthode car toutes les variations candidates n'ont pas été analysées. Cependant, dans les trois races laitières françaises, nous avons sélectionné 482 variants à tester (95 pour la Normande, 152 en Montbéliarde et 235 pour la Holstein). Au final, nous avons mis en évidence un seul polymorphisme candidat avec déficit complet significatif en homozygotes pour l'allèle délétère : le variant g. Chr11 :72399397T>C dans le gène *CAD* ségrégant en race Normande).

Une des difficultés de cette approche réside dans le choix le plus pertinent des variations candidates à tester, de façon à maximiser le nombre de variants réellement responsables d'anomalies. En effet, le taux de succès est assez faible, avec en moyenne 1 à 5 mutations récessives létales portées par individu (MacArthur *et al.*, 2012 ; Charlier *et al.*, 2016).

Deux points ont fortement limité l'efficacité de notre approche : les artéfacts et la fréquence des variants.

a) Maîtriser les artéfacts

Lorsque l'on cherche à prédire les effets des variations rares et délétères, on a tendance à enrichir la sélection en erreurs : fausses variations dues à des erreurs d'alignement des séquences ou des artéfacts créés par le séquençage (MacArthur *et al.*, 2012 ; Sulem *et al.*, 2015).

Comme discuté dans le chapitre 5 nous n'avons pas appliqué assez de filtres de vérification de la qualité de séquençage des variants identifiés puis sélectionnés pour le génotypage et une proportion non négligeable des variants choisis se sont révélés monomorphes. Il aurait été préférable d'appliquer des tests plus stringents sur la couverture et la qualité de la séquence, vérifier manuellement la présence du variant par analyse des fichiers de séquence (fichier bam) à l'aide d'IGV (Thorvaldsdóttir *et al.*, 2013), puis de vérifier l'existence des variants candidats par PCR et séquençage Sanger avant de les proposer sur la puce.

Par ailleurs, nous avons été limités par un autre type d'artéfact dû à l'annotation fonctionnelle insuffisante des variants. Celle-ci dépend en partie des outils bioinformatiques utilisés, mais aussi et surtout du set de transcrits pris en compte pour la réalisation des annotations des génomes (MacCarthy *et al.*, 2014). Dans ce projet, par choix et en fonction du traitement des données de séquence au sein de l'équipe, nous avons privilégié un seul outil de prédiction pour l'annotation des petits SNP et Indels : l'outil Ve!P basé sur les sets de prédiction de gènes Ensembl (MacLaren *et al.*, 2010). Par comparaison manuelle, nous avons identifié des artéfacts dus à la prédiction de gènes situés à des positions différentes (différences entre les prédictions Ensembl et la position sur UCSC comparées avec l'alignement des gènes d'autres espèces). Nous avons également observé des exons absents, ou des incertitudes sur les sites donneurs et accepteurs d'épissage.

La prise en compte des artéfacts d'annotation au début de ma thèse nous a permis d'acquérir plus d'expérience et d'appliquer par la suite des étapes supplémentaires de vérification manuelle de chaque polymorphisme au cours la seconde étude sur les variations plus fréquentes (voir matériel et méthode Michot *et al.*, 2016). Les travaux à partir des données de séquences montrent l'importance de développer des outils et d'utiliser des critères stricts pour détecter et discriminer ces artéfacts des variations réellement existantes.

b) Des variants trop peu fréquents ?

L'étude d'un variant candidat repose sur son génotypage par puce dans une population de grande taille. L'idée est de rechercher l'absence d'homozygotes pour l'allèle alternatif. Si l'allèle a une fréquence f, la proportion attendue d'homozygote est f^2, une valeur très faible si l'allèle est rare. Ainsi, par exemple, pour une fréquence f de 1%, la fréquence d'homozygotes est de 1/10000, et un déficit marqué nécessite plusieurs dizaines de milliers de typages, une quantité que l'on n'observe que dans peu de races. Dans la plupart des races, on ne peut tester que des variants assez fréquents, dont un nombre très limité est sans doute fortement délétère. Dans notre cas, nous aurions pu appliquer un seuil sur la fréquence des allèles délétères dans les bases de données de séquence étudiées, de façon à ne retenir que des polymorphismes que nous pourrions valider statistiquement par génotypage dans nos populations.

II. Perspectives de poursuite des travaux de recherche

A. Suites à court terme des travaux de cette thèse sur les approches inverses

Les différentes analyses menées au cours de ma thèse ont abouti à une grande quantité de données intermédiaires dont seulement une partie a pu être analysée. Nous proposons ici quelques pistes à explorer à partir des résultats de typage obtenus et des polymorphismes dégagés.

1. Suivi des polymorphismes candidats et confirmation des déficits en homozygotes

Un premier travail sera de continuer le suivi des polymorphismes candidats qui sont génotypés sur la puce EuroG10K, en particulier pour ceux dont nous n'avons pas observé d'individus homozygotes. L'accumulation de données de typage permettra de confirmer ces déficits sur un plus grand nombre d'individus de façon à obtenir un résultat significatif d'un point de vue statistique.

L'expérience acquise avec l'haplotype de mort embryonnaire MH1 et l'invalidation d'un variant candidat publié dans l'étude de Fritz *et al.*, 2013 (cf. *Chapitre 4*), nous a montré la nécessité de confirmer le déficit observé sur un très grand nombre d'individus.

Pour les polymorphismes présentant un déficit complet et significatif en homozygotes, tel le variant chr11:72399397T>C dans le gène *CAD* ségrégant en race Normande, il faudra confirmer la mortalité embryonnaire supposée. Une première stratégie serait d'identifier les animaux porteurs de la mutation (ie. directement à partir des données de génotypage, ou par le développement d'un test sur haplotype et l'imputation des génotypes) afin d'étudier la distribution des génotypes chez les veaux nés des accouplements entre porteurs.

Comme pour le gène *PFAS*, l'effet de la mutation du gène *CAD* peut être testé sur différents indicateurs de fertilité calculés à partir des bases de données nationales d'insémination (ie. les taux de non retour à 56 et 90 jours, le taux de conception, l'intervalle entre deux IA ou l'intervalle vêlage-vêlage). Cette seconde stratégie utilisant les données de la base d'inséminations française, permet de caractériser l'effet sur la fertilité et de caractériser le moment de la perte embryonnaire au cours de la gestation (Fritz *et al.*, 2013, Sahana *et al.*, 2013). L'avantage est de pouvoir s'appuyer sur des données déjà disponibles. En revanche, il faut réussir à distinguer un effet significatif, ce qui nécessite de disposer d'un effectif d'accouplements à risque suffisant. Naturellement cet effectif dépend de la fréquence du variant délétère dans la population étudiée.

Enfin, le suivi de gestations résultant d'accouplements à risque ou de transferts d'embryons homozygotes mutés est la stratégie idéale pour caractériser très finement le processus engendrant la perte embryonnaire. Cependant, c'est une procédure lourde et coûteuse qui demande la production puis le transfert d'un nombre suffisant d'embryons. Elle est donc peu généralisable chez le bovin et la confirmation statistique de l'absence d'homozygotes par génotypage large sur la population reste le moyen de validation le plus efficace.

2. Etude de survie et phénotypage des animaux homozygotes mutés

Concernant les polymorphismes pour lesquels sont observés des individus homozygotes pour l'allèle supposé délétère (avec un déficit ou non), nous proposons de mettre en place un suivi de ces individus homozygotes. En effet, si ces animaux apparaissent normaux au moment du génotypage, les symptômes peuvent apparaître plus tardivement et conduire à la mort ou la réforme précoce de l'animal sans que l'on ne soupçonne une anomalie. C'est le cas pour des troubles métaboliques, ou immunitaires qui peuvent être aisément confondus avec des affections d'origine environnementale.

L'apparition des symptômes peut être tardive comme pour l'ataxie progressive en race Charolaise (www.onab.fr) ou encore la cardiomyopathie en Holstein Red (Owczarek-Lipska, 2011). Un suivi des courbes de mortalité des individus homozygotes mutés en comparaison avec les autres génotypes peut être un moyen additionnel de détecter un effet délétère du polymorphisme. Enfin, on peut envisager le phénotypage fin des individus homozygotes pour quelques mutations prioritaires quand le phénotype attendu est bien défini, comme nous l'avons entrepris pour la confirmation de l'effet de la mutation du gène *RP1* et la perte de vision progressive en race Normande.

3. Imputation et analyses d'association sur les caractères de production

Les polymorphismes que nous avons testés sont certainement en grande partie non létaux et n'ont pas d'effet majeur. En revanche, ils pourraient intervenir dans le déterminisme de caractères complexes, ou bien affecter le niveau de performance des animaux porteurs homozygotes ou hétérozygotes (comme par exemple le variant WWP1 p.R844Q augmentant la musculature et diminuant la stature des animaux en Blanc Bleu Belge, Charlier *et al.*, 2016). Un tel effet est difficile à mettre en évidence à partir de quelques cas. Il nécessite de tester l'association du polymorphisme avec les différents caractères mesurés pour la sélection (production, croissance, longévité, fertilité…). Les animaux typés étant jeunes, il faudrait attendre plusieurs mois ou années pour qu'ils réalisent les performances nécessaires. Une alternative est d'imputer les génotypes aux polymorphismes étudiés sur une population plus vieille qui dispose déjà de phénotypes enregistrés. Ces travaux d'imputation sont développés dans l'équipe et les polymorphismes isolés au cours de ma thèse sont en cours d'analyse dans une étude globale.

B. Validation et étude fonctionnelle des mutations candidates

Lorsque c'est possible, la description clinique fine de l'anomalie permet de confirmer l'association de la mutation et son effet macroscopique sur le phénotype de l'individu. Ceci est suffisant chez le bovin pour justifier la mise en place d'une contre-sélection (voir ci-dessous paragraphe IV-A). Dans certains cas, une étude fonctionnelle du gène et de l'effet de la mutation est envisagée, en particulier lorsque la fonction du gène n'est pas ou mal connue. C'est une perspective envisagée en particulier pour la mutation candidate associée à l'épilepsie en race Parthenaise. Comme nous l'avons suggéré, la suite à court terme de ce projet sur l'épilepsie est de confirmer l'effet de la substitution sur l'épissage du gène *MTCL1*.

Parmi les méthodes de validation, la caractérisation de l'ARN, ainsi que de la protéine correspondante par immunohistochimie peut être envisagée. L'intérêt ici est de comprendre les mécanismes moléculaires à l'origine des crises d'épilepsies.

III. Développement de l'utilisation des données de séquence : vers une détection systématique des mutations délétères ?

Avec le développement rapide des techniques de séquençage, la quantité de données de séquence disponibles a triplé au cours de mes trois années de thèse. L'accès à ces données a été mis à profit pour l'étude des anomalies et le développement d'approches de génétique inverse de caractérisation des variants, comme le montrent les publications récentes (Das *et al.*, 2015 ; Boussaha *et al.*, 2016 ; Charlier *et al.*, 2016, Michot *et al.*, 2016). Nous entrons dans l'ère post-génomique où le séquençage des animaux

domestiques va devenir d'usage courant, avec déjà une automatisation des pipelines de détection et d'annotation des petites variations du génome affectant la partie codante des gènes. Quelles sont les perspectives et les enjeux du développement des stratégies de génétique inverse et d'exploitation des données de séquences ?

A. Explorer d'autres régions du génome

Jusqu'à présent, les travaux de génétique inverse se sont concentrés principalement sur les mutations perte de fonction (LOF) et non synonymes (missense) délétères dans les régions codantes. Si l'expérience montre que les mutations responsables d'anomalies génétiques correspondent souvent à ce type de variants, on ne peut exclure les autres variations, du fait de leur grand nombre (plus de 1000 fois supérieur) et de leur annotation très insuffisante.

Les premières cibles dans toute analyse sont généralement les variants de petite taille, SNP ou petits InDels. Les variants structuraux, c'est-à-dire les variants de grande taille (délétions, insertions, duplications, inversions) jouent un rôle tout particulier, car ils couvrent une région chromosomique, potentiellement fonctionnelle, plus grande. Certes, ils sont beaucoup moins nombreux que les variants d'une ou de quelques bases (on observe environ 90% de SNP, 9% de petits indels et 1% de variants structuraux) mais ils ont une probabilité beaucoup plus forte d'avoir un impact fonctionnel, surtout lorsqu'ils affectent tout un gène ou une partie importante de ce gène.

Les exemples chez les bovins ne manquent pas. Nous avons montré dans cette thèse qu'une délétion de 4,8 kb dans le gène *ITGB4* induisait une épidermolyse bulleuse jonctionnelle en race Charolaise. Capitan *et al.* (2012) ont montré l'impact d'une grande délétion incluant le gène *ZEB2* en race charolaise, avec des conséquences sévères sur différentes fonctions. Kadri *et al.* (2014) ont mis en évidence une grande délétion de 660 kb avec un effet très délétère sur la fertilité en race Rouge nordique. Un autre exemple est la délétion du gène *FANCI*, responsable du syndrome Brachyspina en race Holstein (Charlier *et al.*, 2012). Plusieurs équipes ont mis en évidence le déterminisme du syndrome CDH (déficience de cholestérol) en race Holstein, dû à l'insertion d'un retrotransposon dans le gène *APOB* (Menzi *et al.*, 2016), conduisant à sa perte de fonction. Dans la plupart de ces exemples, on note que les variants structuraux entrainent la disparition de parties importantes de gènes, voire de gènes entiers. Le dernier cas est différent, car il correspond à l'insertion d'un élément mobile au milieu d'un gène, altérant sa fonction. Les variants structuraux sont donc des cibles privilégiées de ces recherches d'anomalies à partir des séquences. Parfois, les variants structuraux ne touchent pas des gènes mais plutôt des régions régulatrices. C'est le cas par exemple des deux polymorphismes responsables de l'absence de cornes chez la vache, une insertion et une délétion. Le mécanisme n'est pas connu, mais il est sans doute lié à une interaction avec un lncRNA exprimé dans cette région.

D’un point de vue méthodologique, dans une région spécifiée, un outil de diagnostic particulièrement efficace est *Integrative Genomics Viewer* (IGV, Thorvaldsdóttir *et al*., 2013). A l’échelle du génome entier, d’autres outils plus systématiques sont utilisés. Boussaha *et al.* (2015) présentent ces approches ainsi qu’un premier inventaire des variants structuraux observés dans trois races françaises à partir de données de séquence de génome complet.

La recherche systématique et l’annotation des variants structuraux est donc une priorité pour la caractérisation des anomalies. C’est la voie à explorer si la recherche de variants affectant la partie codante des gènes n’aboutit pas. On ne peut pas non plus exclure des mutations ponctuelles dans les régions régulatrices, comme cela a été montré pour le phénotype culard en race ovine Texel (Clop *et al.*, 2006).

B. Développer l’intégration des données et améliorer l’annotation du génome pour améliorer la compréhension des variations

L’analyse des séquences offre des opportunités considérables, dès lors que l’on sait en extraire toute l’information. Les résultats sont nombreux et difficilement validables de façon exhaustive dans des analyses fonctionnelles de laboratoire. Il est donc essentiel de disposer des méthodes bioinformatiques les plus sûres, évitant aussi bien les faux positifs que les faux négatifs. Les faux positifs s’expliquent par le fait que tous les gènes ne sont pas essentiels pour le développement et aussi en grande partie par les redondances entre voies métaboliques, permettant de maintenir la fonction même quand une des voies est atteinte.

Deux raisons principales expliquent les faux négatifs :

- la non-détection d’une mutation, principalement par manque de couverture de séquence à ce locus, sa présence dans une zone répétée rendant son analyse trop compliquée, ou un trou dans la séquence de référence conduisant à la perte de l’information lors de la phase d’alignement sur la référence des séquences produites ;
- l’absence d’annotation informative d’un variant détecté.

La qualité des résultats dépend de la couverture de séquençage, elle-même très liée au coût. La baisse des coûts permet une meilleure qualité de couverture de séquençage d’un génome. Les nouvelles plateformes de séquençage devraient permettre une augmentation de la couverture par animal, ainsi que l’augmentation de la taille des lectures, ce qui augmente le chevauchement entre elles et la qualité d’alignement.

Des progrès sont nécessaires dans la qualité du génome de référence bovin dont dépendent les alignements et tous les travaux de re-séquençage de génome. De nouvelles versions sont attendues dans un avenir proche, obtenues avec les approches de séquençage les plus récentes.

D'autres progrès sont également nécessaires dans la détection des variants non codants impliqués dans les mécanismes de régulation des gènes. Cela passe par l'amélioration de l'annotation du génome. C'est l'objectif par exemple du projet FR-AgENCODE, dont le but est d'améliorer l'annotation des génomes des espèces domestiques comme cela a été fait dans le cadre du projet ENCODE humain. Plus particulièrement, ce projet vise à mesurer l'effet des polymorphismes sur l'expression, à identifier les réseaux de gènes co-régulés, à identifier les régions régulatrices et les interactions avec les facteurs de transcription. A terme, l'intégration de l'ensemble des données est une voie envisagée pour identifier de façon plus fiable les variants délétères dans les anomalies qu'elles soient monogéniques ou à déterminisme plus complexe (Jiang *et al.*, 2015).

C. Vers le séquençage systématique des animaux reproducteurs

Jusqu'ici, la priorité de re-séquençage a été donnée à deux types d'animaux :

- d'une part, aux taureaux fondateurs de races et contributeurs majoritaires, dans le but de capter la majorité des polymorphismes ségrégant dans les races et de permettre l'imputation des génotypes, un phasage efficace et la détection des mutations causales pour expliquant les variations des caractères (Daetwyler *et al.*, 2014) ;
- d'autre part, les animaux avec des phénotypes particuliers, comme les anomalies.

Dans le cadre de cette thèse, les données étaient constituées des séquences de 429 animaux appartenant à 4 races différentes. Aujourd'hui, le Run 4 du consortium « 1000 bull genomes » compte trois fois plus d'individus séquencés pour 33 races différentes.

Poussé par les besoins en génétique humaine, le développement de techniques et des plateformes de séquençage est toujours plus efficace. Une évolution récente chez Illumina est le séquenceur haut-débit HISEQ X qui laisse entrevoir la possibilité de séquencer des génomes humains à une couverture de 30x pour un coût de 1000 dollars (Check Haden, 2014). Cette technologie a été ouverte fin 2015 aux espèces domestiques et donc aux bovins, augmentant à nouveau la capacité de re-séquençage de génomes (Business Wire, 2015).

A court ou moyen terme, cette technique et d'autres en cours de développement offriront la possibilité technique et financière de réaliser un séquençage de génome d'ordre populationnel dans les espèces domestiques. Chez le bovin, on envisage par exemple le séquençage de chaque taureau d'insémination mis en production. Dans ce contexte, le développement de nouvelles approches d'exploitation des données de séquence comme ce fut le cas au cours de ma thèse prend un sens particulier. L'objectif sera d'effectuer une recherche systématique des variants délétères chez chaque nouvel individu reproducteur. En particulier, la détection précoce de toute neo-mutation potentiellement très délétère chez un individu. Ceci sera possible en développant notre capacité de prédiction des effets positifs ou négatifs d'une mutation sur un gène d'intérêt comme indiqué précédemment. Toute mutation ne pourra être testée.

Dans le cas des mutations récessives, l'objectif serait plutôt de repérer les facteurs de risques et limiter l'émergence de nouvelles anomalies en ajustant la gestion des reproducteurs (puisqu'il est impossible de les supprimer tous, voir § IV-A2). Dans une optique plus large, le séquençage de toute la population de reproducteurs permettra aussi de détecter de nouveaux variants d'intérêt pour les caractères de production, reproduction ou de santé.

IV. Intérêt de l'étude des anomalies génétique chez le bovin

A. Gestion des anomalies génétiques dans les programmes de sélection

L'objectif principal de l'étude des anomalies génétiques dans les espèces de rente est avant tout l'identification de la mutation causale et la mise au point de tests de dépistage afin de permettre leur gestion dans la population concernée.

1. Mise en place de tests génétiques

Deux types de tests génétiques peuvent être développés : des tests directs, consistant à génotyper la mutation causale, et des tests indirects, consistant à génotyper un ou plusieurs variants (alors réunis en haplotypes) dont certains allèles présentent un fort déséquilibre de liaison avec la mutation causale. En général, on a recours à ces tests indirects, pour commencer à gérer les anomalies génétiques avant que la mutation causale ne soit identifiée. Lorsque la mutation est connue, l'utilisation de tests sur haplotype permet également, par imputation, de prédire le statut d'animaux déjà génotypés dans le passé et ainsi de récupérer des informations précieuses à l'échelle raciale. Toutefois, ces tests ne sont pas fiables à 100% et le risque de faux positifs et de faux négatifs croit à mesure que le déséquilibre de liaison diminue. C'est notamment le cas pour les mutations anciennes pour lesquelles le segment IBD autour de la mutation est de très petite taille et ne permet pas d'identifier un haplotype unique associé à la mutation (par exemple pour la mutation dans le gène *RP1*), ou encore pour les mutations récentes qui se sont produites sur un haplotype fréquent dans la population et dont la version ancestrale (ie. sans la mutation) perdure dans la population.

Au-delà du fait que la génétique inverse permet d'identifier des anomalies génétiques passées inaperçues jusqu'à présent, cette stratégie présente deux avantages majeurs : l'ensemble des candidats à la sélection génomique est génotypé pour la mutation causale lors de la phase de test, et le test de diagnostic est opérationnel sans délai dès que l'effet de la mutation est validé. Ce transfert rapide des résultats de la recherche au « terrain », sans coût supplémentaire, participe à l'augmentation de l'efficacité de la gestion des anomalies dans les schémas de sélection.

2. Evolution des stratégies de gestion

Lorsque les anomalies connues étaient peu nombreuses, elles étaient gérées par une procréation d'animaux non porteurs et l'élimination progressive des animaux porteurs, en général sur la voie mâle, indépendamment du reste de l'objectif de sélection. Avec les méthodes les plus récentes, l'efficacité de détection des anomalies s'accroit sensiblement, de sorte que le nombre d'anomalies à gérer simultanément s'accroit également. Il devient difficile de sélectionner des reproducteurs exempts de toute anomalie et il est nécessaire de revoir les procédures de sélection. La méthode recommandée vise à sélectionner sur la base d'un indice de sélection incluant les anomalies. La valeur génétique d'un reproducteur pour une anomalie donnée s'exprime simplement en fonction de la fréquence allélique et du coût économique imputable à chacun des génotypes. Cette valeur génétique s'intègre ensuite dans l'objectif de sélection. Par ailleurs, en attendant l'éradication de l'anomalie, il convient de gérer les accouplements pour limiter au maximum le risque de procréation de produits atteints. Cette stratégie, dans tous les cas, repose sur un génotypage à grande échelle, permettant de connaitre le statut des reproducteurs et d'identifier les accouplements à risque.

B. Connaissance du génome : le bovin comme modèle

Comme le montre l'exemple de l'épilepsie idiopathique récessive en race Parthenaise, les recherches portant sur les animaux d'élevage peuvent parfois contribuer à l'amélioration de l'annotation fonctionnelle des génomes à travers la caractérisation de l'effet de mutations affectant des gènes dont la fonction n'a pas encore été étudiée chez les espèces modèles. Par rapport aux espèces modèles, les espèces d'élevage comme le bovin présentent d'autres avantages qui pourraient leur permettre de devenir des animaux modèles à part entière à l'ère post-génomique. Ces derniers sont détaillés dans l'article de Bourneuf *et al.,* (soumis) dont je suis co-auteure. On retiendra notamment:

- la taille importante des familles qui permet (i) de recruter un nombre suffisant de cas et (ii) de cartographier des loci modificateurs ;
- la taille réduite de l'effectif efficace des races combinée aux importants efforts de séquençage qui permettent (i) de capter presque l'intégralité de la variabilité génétique d'une race et (ii) d'identifier efficacement des mutations *de novo* ; la richesse des bases de données de génotypage sur puce à SNP, de pedigree et d'insémination qui permet d'identifier des accouplements à risque et de phénotyper les animaux qui en sont issus ;
- et enfin l'existence de bases de données de phénotypes collectés en routine dans le cadre des évaluations génétiques pour une quarantaine de caractères en moyenne aujourd'hui et un nombre bien plus important à l'avenir.

L’exploitation de ces phénotypes représente un enjeu important de l’approche de génétique inverse qui n’a pu être mené à bien dans le cadre de cette thèse faute de temps. Ce sera l’un des objectifs principaux de mon équipe d’accueil dans les prochains mois.

Références bibliographiques

Adams HA, Sonstegard TS, VanRaden PM, Null DJ, Van Tassell CP, Larkin DM et Lewin HA. 2016. Identification of a nonsense mutation in APAF1 that is likely causal for a decrease in reproductive efficiency in Holstein dairy cattle. J. Dairy Sci. 99:6693-6701.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS et Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods. 7:248-249.

Agerholm JS, Basse A, Christensen K. 1993. Investigations on the occurrence of hereditary diseases in the Danish cattle population 1989-1991. Acta Vet Scand. 34:245-53.

Agerholm JS, Bendixen C, Andersen O et Arnbjerg J. 2001. Complex vertebral malformation in Holstein calves. J Vet Diagn Invest. 13, 283-289.

Agerholm JS, Arnbjerg J et Andersen O. 2004. Familial Chondrodysplasia in Holstein Calves. J Vet Diagn Invest. 16:293-298.

Agreste. Infos rapides ─ Animaux de boucherie ─ Bovins ─ cheptel ─ avril 2016 - n°1/2. http://agreste.agriculture.gouv.fr/IMG/pdf/conjinforap201604bvfr.pdf.

Akteson F, Heman W, Ibsen L et Eldridge F. 1944. Inheritance of an epileptic type character in brown Swiss cattle. J Hered. 35:45-48.

Andersson L. 2001. Genetic dissection of phenotypic diversity in farm animals. Nat Rev Genet. 2:130-138.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA et Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 12:745-755.

Benit P, Beugnot R, Chretien D, Giurgea I, De Lonlay-Debeney P, Issartel JP, Corral-Debrinski, M., Kerscher S, Rustin P, Rotig A, Munnich A. 2003. Mutant NDUFV2 subunit of mitochondrial complex I causes early onset hypertrophic cardiomyopathy and encephalopathy. Hum. Mutat. 21: 582-586.

Berendt M, Farquhar RG, Mandigers PJ, Pakozdy A, Bhatti SF, DeRisio L, Fischer A, Long S, Matiasek, K, Munana, K, Patterson EE, Penderis J, Platt S, Podell M, Potschka H, Pumarola MB, Rusbridge C, Stein VM, Tipold A et Volk HA. 2015. International veterinary epilepsy task force consensus report on epilepsy definition, classification and terminology in companion animals. BMC Vet Res 11:182.

Berg AT, Berkovic SF, Brodie MJ, Buchhalter J, Cross JH, van Emde Boas W, Engel J, French J, Glauser TA, Mathern GW, Moshe SL, Nordli D, Plouin P et Scheffer IE. 2010. Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009. Epilepsia. 51:676-85.

Boichard D, Coquereau JA, Amigues Y et Le Mezec P. 1994. Etude de l'anomalie génétique Blad chez les bovins Holstein. 1ères Renc. Rech. Ruminants, 1-2 Décembre 1994, Paris, 1:257-260.

Boichard D, Maignel L et Verrier E. 1996. Analyse généalogique des races bovines laitières françaises. INRA Prod. Anim., 9:323-335.

Boichard D, Maignel L, et Verrier E. 1997.The value of using probabilities of gene origin to measure genetic variability in a population. Genet. Sel. Evol., 29:5-23.

Boichard D. 2002. Pedig : a fortran package for pedigree analysis suited to large populations. In 7th World Congress on Genetics Applied to Livestock Production, Montpellier, 19-23 août 2002, paper 28-13.

Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, Gietzen KJ, Hayes BJ, Lawley CT, Sonstegard TS, Van Tassell CP, Vanraden PM, Viaud K et Wiggans GR. 2012a . Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLoS ONE, 7: e34130.

Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, Druet T, Genestout L, Colleau JJ, Journaux L, Ducrocq V et Fritz S. 2012b. Genomic Selection in French Dairy Cattle. Anim Prod Sci. 52:115–120.

Bouquet A, Renand G, et Phocas F. 2009. Evolution de la diversité génétique des populations françaises de bovins allaitants spécialisés de 1979 à 2008. Inra Prod. Anim. 22:317-330.

Bourneuf E., Otz P., Pausch H., Jagannathan V., Michot P., Grohs C., Piton G., Ammermüller S., Deloche M.C., Fritz S., Leclerc H., Péchoux C., Boukadiri A., Saintilan R., Créchet F., Mosca M., Segelke D., Guillaume F., Bouet S., Baur A., Vasilescu A., Genestout L., Thomas A., Allais-Bonnet A., Rocha D., Colle M.A., Klopp C., Esquerré D., Wurmser C., Flisikowski K., Schwarzenbacher H., Burgstaller J., Brügmann M., Dietschi E., Huth N., Freick M., Barbey S., Fayolle G., Danchin-Burge C., Schibler L., Bed'hom B., Hayes B.J., Daetwyler H.D., Fries R., Boichard D., Pin D., Drögemüller C., Capitan A. 2016. Rapid Discovery of *De Novo* Deleterious Mutations in Cattle Using Genome Sequence Data: Enhancing the Value of Farm Animals as Model Species. Soumis à Scientific Reports.

Boussaha M, Esquerré D, Barbieri J, Djari A, Pinton A, Letaief R, Salin G, Escudié F, Roulet A, Fritz S, Samson F, Grohs C, Bernard M, Klopp C, Boichard D et Rocha D. 2015. Genome-Wide Study of Structural Variants in Bovine Holstein, Montbéliarde and Normande Dairy Breeds. PLoS ONE. 10:e0135931.

Boussaha M, Michot P, Letaief R, Hoze C, Fritz S, Grohs C, Esquerre D, Duchesne A, Philippe R, Blanquet V, Phocas F, Floriot S, Rocha D, Klopp C, Capitan A et Boichard D. 2016. Construction of a large collection of small genome variations in French dairy and beef breeds using whole genome sequences. Genet Sel Evol. 48:87.

Bronwen LA, Ayling S, Barrell D,Clarke L, Curwen V, Fairley S, Banet JF, Billis K, Garcín Girón C, Hourlier T, Howe K, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel JH, White S, Zadissa A, Flicek P et Searle SMJ. 2016. The Ensembl gene annotation system. Database. 2016: baw093

Buisson NB. Epilepsies et malformations du développement cortical. Rédigé par la LFCE Décembre 2013, modifié Septembre 2014 ; http://www.lfce.fr/Epilepsies-et-malformations-du-developpement-cortical_a341.html.

Business Wire. 2015. Illumina Expands Use of HiSeq X™ Sequencing System to Include Non-Human Species, http://www.businesswire.com/news/home/20151006005360/en, consulté le 05/10/2016.

Capitan A, Allais-Bonnet A, Pinton A, Marquant-Le Guienne B, Le Bourhis D, Grohs C, Bouet S, Clément L, Salas-Cortes L, Venot E, Chaffaux S, Weiss B, Delpeuch A, Noé G, Rossignol MN, Barbey S, Dozias D, Cobo E, Barasc H, Auguste A, Pannetier M, Deloche MC, Lhuilier E, Bouchez O, Esquerré D, Salin G, Klopp C, Donnadieu C, Chantry-Darmon C, Hayes H, Gallard Y, Ponsart C, Boichard D, Pailhoux E. 2012. A 3.7 Mb deletion encompassing ZEB2 causes a novel polled and multisystemic syndrome in the progeny of a somatic mosaic bull. PLoS ONE. 7:e49084.

Carletti F, Sardo P, Gambino G, Liu XA, Ferraro G et Rizzo V. 2016. Hippocampal Hyperexcitability is Modulated by Microtubule-Active Agent: Evidence from In Vivo and In Vitro Epilepsy Models in the Rat. Front Cell Neurosci. 10:29.

Cassou R. 1968. 6th Inter. Congr. Animal Reprod. and Artif. Insem. Paris. II: 1009-1012.

Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, Carta E, Dardano S, Dive M, Fasquelle C, Frennet JC, Hanset R, Hubin X, Jorgensen C, Karim L, Kent M, Harvey K, Pearce BR, Simon P, Tama N, Nie H, Vandeputte S, Lien S, Longeri M, Fredholm M, Harvey RJ et Georges M. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet. 40:449–54.

Charlier C, Agerholm JS, Coppieters W, Karlskov-Mortensen P, Li W, de Jong G, Fasquelle C, Karim L, Cirera S, Cambisano N, Ahariz N, Mullaart E, Georges M et Fredholm M. 2012. A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina. PLoS ONE. 7: e43085.

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, Davis S, Druet T, Faux P, Guillaume F, Karim L, Keehan M, Kadri NK, Tamma N, Spelman R et Georges M. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Res. 26:1333-1341.

Check Hayden E. 2014. Is the $1,000 genome for real? Nature News, doi:10.1038/nature.2014.14530.

Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibé B, Bouix J, Caiment F, Elsen JM, Eychenne F, Larzul C, Laville E, Meish F, Milenkovic D, Tobin J, Charlier C et Georges M. 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet. 38:813-818.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C, Esquerre D, Bouchez O, Rossignol MN, Klopp C, Rocha D, Fritz S,

Eggen A, Bowman PJ, Coote D, Chamberlain AJ, Anderson C, VanTassell CP, Hulsegge I, Goddard ME, Guldbrandtsen B, Lund MS, Veerkamp RF, Boichard DA, Fries R et Hayes BJ 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 46: 858-865.

Danchin-Burge C, Leroy G, Brochard M, Moureaux S, Verrier E. 2012. Evolution of the genetic variability of eight French dairy cattle breeds assessed by pedigree analysis. J Anim Breed Genet. 129:206-217.

Danchin-Burge C, Verrier E, Laloë D, Saintilan R, et Leroy G. 2014. An Observatory of the Genetic Variability of Ruminants and Equids breeds. Proceedings, 10th World Congress of Genetics Applied to Livestock Production, Vancouver.

D'Angelo A, Bellino C, Bertone I, Cagnotti G, Iulini B, Miniscalco B, Casalone C, Gianella P et Cagnasso A. 2015. Seizure disorders in 43 cattle. J Vet Intern Med. 29:967-971.

Das A, Panitz F, Gregersen VR, Bendixen C et Holm LE. 2015. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics. 16: 1043

Divina P, Kvitkovicova A, Buratti E et Vorechovsky I. 2009. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. Eur J Hum Genet. 17:759-765

Douaud M, Feve K, Pituello F, Gourichon D. Boitard S, Leguern E, Coquerelle G, Vieaud A, Batini C, Naquet R, Vignal A, Tixier-Boichard M, et Pitel F. 2011. Epilepsy Caused by an Abnormal Alternative Splicing with Dosage Effect of the SV2A Gene in a Chicken Model. PLoS ONE. 6: e26932.

Drewes G, Ebneth A, Preuss U, Mandelkow EV et Mandelkow E. 1997. MARK, a Novel Family of Protein Kinases That Phosphorylate Microtubule-Associated Proteins and Trigger Microtubule Disruption. Cell. 89:297-308.

Drögemüller C., Tetens J., Sigurdsson S., Gentile A., Testoni S., Lindblad-Toh K. et Leeb T. 2010. Identification of the bovine Arachnomelia mutation by massively parallel sequencing implicates sulfite oxidase (SUOX) in bone development. PLoS Genet. 6:e1001079.

Dubey J, Ratnakaran N et Koushika SP. 2015. Neurodegeneration and microtubule dynamics: death by a thousand cuts. Front Cell Neurosci. 9 :343.

Ducos A, Eggen A, Darré R et Boichard D. 2003. Un projet d'épidémiosurveillance pour les anomalies héréditaires bovines. Bull GTV, 18, Dossier spécial "Les anomalies héréditaires des bovins", 41-45.

Elsik CG, *et al* (The Bovine Genome Sequencing and Analysis Consortium). 2009. The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. Science. 5926: 522–528.

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet

D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J et Chee MS. 2003. Highly parallel snp genotyping. Cold Spring Harb Symp Quant Biol. 68:69-78.

Fasquelle C, Sartelet A, Li W, Dive M, Tamma N, Michaux C, Druet T, Huijbers IJ, Isacke CM, Coppieters W, Georges M et Charlier C. 2009. Balancing selection of a frame-shift mutation in the MRC2 gene accounts for the outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. PLoS Genet. 5:e1000666.

FGE. 2016. Chiffres clés 2015. Accessible sur : http://fr.france-genetique-elevage.org/IMG/pdf/chiffre_cles_2015_.pdf.

Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, Engel J, Forsgren L, French JA, Glynn M, Hesdorffer DC, Lee BI, Mathern GW, Moshe SL, Perucca E, Scheffer IE, Tomson T, Watanabe M et Wiebe S. 2014. ILAE official report: a practical clinical definition of epilepsy. Epilepsia. 55:475-82.

Floriot S, Vesque C, Rodriguez S, Bourgain-Guglielmetti F, Karaiskou A Gautier M, Duchesne A, Barbey S, Fritz S, Vasilescu A, Bertaud M, Moudjou M, Halliez S, Cormier-Daire V, Hokayem JE, Nigg EA, Manciaux L, Guatteo R, Cesbron N, Toutirais G, Eggen A, Schneider-Maunoury S, Boichard D, Sobczak-Thépot J, et Schibler L. 2015. C-Nap1 mutation affects centriole cohesion and is associated with a Seckel-like syndrome in cattle Nat Commun.6:6894.

Fournier D, Keppie N et Simko E. 2004. Bovine familial convulsions and ataxia in Saskatchewan and Alberta. Can Vet J. 45:845-8.

Fritz S, Capitan A, Djari A, Rodriguez SC, Barbat A, Baur A, Grohs C, Weiss B, Boussaha M, Esquerré D, Klopp C, Rocha D et Boichard D. 2013. Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2. PLoS ONE. 8: e65550.

George AL. 2004. Inherited Channelopathies Associated with Epilepsy. Epilepsy Currents. 4:65–70.

Gonzaga-Jauregui C, Gamble CN, Yuan B, Penney S, Jhangiani S, Muzny DM, Gibbs RA, Lupski JR et Hecht JT. 2015. Mutations in COL27A1 cause Steel syndrome and suggest a founder mutation effect in the Puerto Rican population. Eur J of Hum Genet. 23:342-346.

Grande-García A, Lallous N, Díaz-Tejada C, et Ramón-Maiques S. 2014. Structure, functional characterization, and evolution of the dihydroorotase domain of human CAD.Structure. 22:185-98.

Grobet L, Poncelet D, Royo LJ, Brouwers B, Pirottin D, Michaux C, Ménissier F, Zanotti M, Dunner S et Georges M. 1998. Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. Mamm Genome. 9:210-213

Guaguere E, Berg K, Degorce-Rubiales F, Spadafora A et Meneguzzi G. 2004. Junctional epidermolysis bullosa in a Charolais calf with deficient expression of integrin α6ß4. Vet Dermatol.15:28.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG et Chee MS. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. Nat. Genet. 37:549-554.

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mullaart E, Coppieters W et Georges M. 2016. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions. http://biorxiv.org/content/early/2016/10/09/079863.

Helbig I. 2015. Genetic Causes of Generalized Epilepsies. Semin Neurol. 35:288-292.

Hirose S. 2014. Mutant GABA(A) receptor subunits in genetic (idiopathic) epilepsy. Prog Brain Res. 213:55-85.

Jallon P et Latour P. 2005. Epidemiology of idiopathic generalized epilepsies. Epilepsia 46 (Suppl 9):10-14.

Jiang R, Wu M et Li L. 2015. Pinpointing disease genes through phenomic and genomic data fusion. BMC Genomics. 16 (Suppl 2):S3.

Kadri NK, Sahana G, Charlier C, Iso-Touru T, Guldbrandtsen B, Karim L, Nielsen US, Panitz F, Aamand GP, Schulman N, Georges M, Vilkki J, Lund MS et Druet T. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. PLoS Genet. 10:e1004049.

Kheradmand Kia S, Verbeek E, Engelen E, Schot R, Poot RA, de Coo IFM, Lequin MH, Poulton CJ, Pourfarzad F, Grosveld FG, Brehm A, de Wit MCY, Oegema R, Dobyns WB, Verheijen FW et Mancini GMS. 2012. RTTN mutations link primary cilia function to organization of the human cerebral cortex. Am J Hum Genet. 91: 533-540.

Koskinen LL, Seppala EH, Belanger JM, Arumilli M, Hakosalo O, Jokinen P, Nevalainen EM, Viitmaa R, Jokinen TS, Oberbauer AM, et Lohi H. 2015. Identification of a common risk haplotype for canine idiopathic epilepsy in the ADAM23 gene. BMC Genomics. 16:465.

Kumar P, Henikoff S et Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 4:1073-1081.

Lacombe VA, Mayes M, Mosseri S, Reed SM, Fenner WR, et Ou HT. 2012. Epilepsy in horses: aetiological classification and predictive factors. Equine Vet J. 44: 646-651.

Lander ES et Botstein D. 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 236:1567-1570.

Lee K, Nguyen DT, Choi M, Cha SY, Kim JH, Dadi H, Seo HG, Seo K, Chun T et Park C. 2013. Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant. BMC Genomics. 14:596.

Lenffer J, Nicholas FW, Castle K, Rao A, Gregory S, Poidinger M, Mailman MD et Ranganathan S. 2006. OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. Nucleic Acids Res. 34:D599-601.

Li H et Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R et 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Li W, Sartelet A, Tamma N, Coppieters W, Georges M, Charlier C. 2016. Reverse genetic screen for loss-of-function mutations uncovers a frameshifting deletion in the melanophilin gene accountable for a distinctive coat color in Belgian Blue cattle. Anim Genet. 47:110-113.

Lohi H, Young EJ, Fitzmaurice SN, Rusbridge C, Chan EM, Vervoort M, Turnbull J, Zhao XC, Ianzano L, Paterson AD, Sutter NB, Ostrander EA, Andre C, Shelton GD, Ackerley CA, Scherer SW et Minassian BA. 2005. Expanded repeat in canine epilepsy. Science. 307:81.

Lynch M. 2016. Mutation and Human Exceptionalism: Our Future Genetic Load. Genetics. 202:869-875

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, 1000 Genomes Project Consortium, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB et Tyler-Smith C. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 335:823-828.

Macdonald RL, Kang JQ, et Gallagher MJ. 2010. Mutations in GABAA receptor subunits associated with genetic epilepsies. J Physiol. 588:1861-1869.

Mattalia S, Barbat A, Danchin-Burge C, Brochard M, Le Mezec P, Minery S, Jansen G, Van Doormaal B et Verrier E. 2006. La variabilité génétique des huit principales races bovines laitières françaises : quelles évolutions, quelles comparaisons internationales ? Renc. Rech. Ruminants, 13 : 239-246.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS et Van Tassell CP. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE. 4:e5350.

McClure MC, Bickhart D, Null D, VanRaden P, Xu LY, Wiggans G, Liu G, Schroeder S, Glasscock J, Armstrong J, Cole JB, Van Tassell CP et Sonstegard TS. 2014. Bovine Exome Sequence Analysis and Targeted SNP Genotyping of Recessive Fertility Defects BH1, HH2, and HH3 Reveal a Putative Causative Mutation in SMC2 for HH3. PLoS ONE 9:e92769.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et DePristo MA. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing nextgeneration DNA sequencing data. Genome Res. 20: 1297–1303.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17:122.

Menzi F, Besuchet-Schmutz N, Fragnière M, Hofstetter S, Jagannathan V, Mock T, Raemy A, Studer E, Mehinagic K, Regenscheit N, Meylan M, Schmitz-Hsu F et Drögemüller C. 2016. A transposable element insertion in APOB causes cholesterol deficiency in Holstein cattle. Anim Genet. 47:253-257.

Michot P, Chahory S, Marete A, Grohs C, Dagios D, Donzel E, Aboukadiri A, Deloche MC, Allais-Bonnet A, Chambrial M, Barbey S, Genestout L, Boussaha M, Danchin-Burge C, Fritz S, Boichard D et Capitan A. 2016. A reverse genetic approach identifies an ancestral frameshift mutation in RP1 causing recessive progressive retinal degeneration in European cattle breeds. Genet Sel Evol. 48:56.

Moureaux S, Boichard D et Verrier E. 2000. Utilisation de l'information généalogique pour l'estimation de la variabilité génétique de 8 races bovines laitières françaises d'extension nationale ou régionale. Renc Rech Ruminants, 7:149-152.

Ng BG, Wolfe LA, Ichikawa M, Markello T, He M, Tifft CJ, Gahl WA et Freeze HH. 2015. Biallelic mutations in CAD impair *de novo* pyrimidine biosynthesis and decrease glycosylation precursors. Hum Molec Genet. 24: 3050-3057.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 42:30-35.

Nielsen R, Paul JS, Albrechtsen A et Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics. 12: 443-451.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, et Pruitt KD. 2016. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44(D1):D733-45.

OMS (Organisation Mondiale de la Santé). Épilepsie Aide-mémoire N°999. Février 2016. http://www.who.int/mediacentre/factsheets/fs999/fr, consulté le 05/06/2016.

Owczarek-Lipska M, Plattet P, Zipperle L, Drögemüller C, Posthaus H, Dolf G et Braunschweig MH. 2010. A nonsense mutation in the optic atrophy 3 gene (OPA3) causes dilated cardiomyopathy in Red Holstein cattle. Genomics. 97: 51-57.

Park L. 2011. Effective population size of current human population. Genet Res (Camb). 93:105-114.

Pausch H., Schwarzenbacher H., Burgstaller J., Flisikowski K., Wurmser C., Jansen S., Jung S., Schnieke A., Wittek T., et Fries R. 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. BMC Genomics. 16: 312.

Peters M, Reber I, Jagannathan V, Raddatz B, Wohlsein P and Drögemüller C. 2015. DNA-based diagnosis of rare diseases in veterinary medicine: a 4.4 kb deletion of ITGB4 is associated with epidermolysis bullosa in Charolais cattle. BMC Vet Res. 11:48.

Rahman S, Footitt EJ, Varadkar S et Clayton PT. 2013. Inborn errors of metabolism causing epilepsy. Dev Med Child Neurol. 55:23-36.

Ribeca C, Bonfatti V, Cecchinato A, Albera A, Gallo L et Carnier P. 2013. Effect of polymorphisms in candidate genes on carcass and meat quality traits in double muscled Piemontese cattle. Meat Sci. 96:1376-1383.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G et Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol. 29:24-26.

Rozen S et Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 132:365-86.

Sahana G, Sander Nielsen U, Pedersen Aamand G, Lund MS et Guldbrandtsen B. 2013. Novel Harmful Recessive Haplotypes Identified for Fertility Traits in Nordic Holstein Cattle. PLoS One. 8: e82909.

Sakakibara A, Ando R, Sapir T et Tanaka T. 2013. Microtubule dynamics in neuronal morphogenesis. Open Biol.3:130061.

Sargolzaei M, Chesnais JP, and Schenkel FS. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 15:478.

Sartelet A, Druet T, Michaux C, Fasquelle C, Géron S, Tamma N, Zhang Z, Coppieters W, Georges M et Charlier C. 2012. A splice site variant in the bovine rnf11 gene compromises growth and regulation of the inflammatory response. PLoS Genet. 8:e1002581.

Sartelet A, Stauber T, Coppieters W, Ludwig CF, Fasquelle C, Druet T, Zhang Z, Ahariz N, Cambisano N, Jentsch TJ et Charlier C. 2014. A missense mutation accelerating the gating of the lysosomal cl−/h+-exchanger clc-7/ostm1 causes osteopetrosis with gingival hamartomas in cattle. Dis Model Mech. 7: 119–128.

Sato Y, Akitsu M, Amano Y, Yamashita K, Ide M, Shimada K, Yamashita A, Hirano H, Arakawa N, Maki T, Hayashi I, Ohno S et Suzuki A. 2013. The novel PAR-1-binding protein MTCL1 has crucial roles in organizing microtubules in polarizing epithelial cells. J Cell Sci. 126: 4671-83.

Sato Y, Hayashi K, Amano Y, Takahashi M, Yonemura S, Hayashi I, Hirose H, Ohno S et Suzuki A. 2014. MTCL1 crosslinks and stabilizes non-centrosomal microtubules on the Golgi membrane. Nat Comm. 5:5266.

Schena M, Shalon D, Davis RW et Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science. 5235:467-470.

Schutz E, Wehrhahn C, Wanjek M, Bortfeld R, Wemheuer WE, Beck J et Brenig B. 2016. The Holstein Friesian Lethal Haplotype 5 (HH5) Results from a Complete Deletion of TBF1M and Cholesterol Deficiency (CDH) from an ERV-(LTR) Insertion into the Coding Region of APOB. PloS ONE. 11: e0154602.

Schwarzenbacher H, Burgstaller J, Seefried FR, Wurmser C, Hilbe M, Jung S, Fuerst C, Dinhopl N, Weissenbock H, Fuerst-Waltl B, Dolezal M, Winkler R, Grueter O, Bleul U, Wittek T, Fries R et Pausch H. 2016. A missense mutation in TUBD1 is associated with high juvenile mortality in Braunvieh and Fleckvieh cattle. BMC Genomics. 17: 400.

Schwenger B, Schober S et Simon D. 1993. DUMPS Cattle Carry a Point Mutation in the Uridine Monophosphate Synthase Gene. Genomics. 16: 241-244.

Seppala EH, Jokinen TS, Fukata M, Fukata Y, Webster MT, Karlsson EK, Kilpinen SK, Steffen F, Dietschi E, Leeb T, Eklund R, Zhao X, Rilstone JJ, Lindblad-Toh K, Minassian BA et Lohi H. 2011. LGI2 truncation causes a remitting focal epilepsy in dogs. PLoS Genet. 7:e1002194.

Shamseldin H, Alazami AM, Manning M, Hashem A, Caluseiu O, Tabarki B, Esplin E, Schelley S, Innes AM, Parboosingh JS, Lamont R, Care4Rare Canada Consortium, Majewski J, Bernier FP et Alkuraya FS. 2015. RTTN mutations cause primary microcephaly and primordial dwarfism in humans. Am J Hum Genet 97: 862-868.

Shorvon SD. 2011. The etiologic classification of epilepsy. Epilepsia. 52: 1052-7.

Shuster DE, Kehrli ME Jr, Ackermann MR et Gilbert RO. 1992. Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. Proc Natl Acad Sci USA. 89: 9225–9229.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD et Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 7:539.

Sonstegard TS, Cole JB, VanRaden PM, Van Tassell CP, Null DJ, Schroeder SG, Bickhart D, McClure MC. 2013. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. PLoS ONE. 8: e54872.

Sorge G et Sorge A. 2010. Epilepsy and chromosomal abnormalities. Italian Journal of Pediatrics. 36: 36.

Steemers FJ, Chang W, Lee G, Barker DL, Shen R et Gunderson KL. 2006. Whole-genome genotyping with the single-base extension assay. Nat Methods. 3:31-33.

Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, Hjartarson E, Sigurdsson GT, Jonasdottir A, Jonasdottir A, Sigurdsson A, Magnusson OT, Kong A, Helgason A, Holm H, Thorsteinsdottir U, Masson G, Gudbjartsson DF et Stefansson K. 2015. Identification of a large set of rare complete human knockouts. Nat Genet. 47:448-452.

Tevosian SG, Deconinck AE, Tanaka M, Schinke M, Litovsky SH, Izumo S, Fujiwara Y, Orkin SH. 2000.FOG-2, a cofactor for GATA transcription factors, is essential for heart morphogenesis and development of coronary vessels from epicardium. Cell.101:729-739.

Thomsen B, Horn P, Panitz F, Bendixen E, Petersen AH, Holm LE, Nielsen VH, Agerholm JS, Arnbjerg J et Bendixen C. 2006. A missense mutation in the bovine SLC35A3 gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation. Genome.Research 16: 97-105.

Thomsen B, Nissen PH, Agerholm JS et Bendixen C. 2010. Congenital bovine spinal dysmyelination is caused by a Faux-sens mutation in the SPAST gene. Neurogenetics 11: 175-183.

Thorvaldsdóttir H, Robinson JT et Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14:178-192.

VanRaden PM, Olson KM, Null DJ et Hutchison JL. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J Dairy Sci. 94:6153–6161.

Venhoranta H, Pausch H, Flisikowski K, Wurmser C, Taponen J, Rautala H, Kind A, Schnieke A, Fries R, Lohi H et Andersson M. 2014. In frame exon skipping in UBE3B is associated with developmental disorders and increased mortality in cattle. BMC Genomics. 15:890.

Verrier E, Rognon X, Laloë D et de Rochambeau H. 2005. Ethnozootechnie. 76: 67-82.

Vignal A, Milan D, SanCristobal M et Eggen A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol. 34:275.

Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, Van Broeckhoven C et De Rijk P. 2005. novoSNP, a novel computational tool for sequence variation discovery. Genome Res. 15: 436-42.

Xu X, Hu Y, Xiong Y, Li Z, Wang W, Du C, Yang Y, Zhang Y, Xiao F, et Wang X. 2016. Association of Microtubule Dynamics with Chronic Epilepsy. Mol Neurobiol 53, 5013-24.

Ye K, Schulz MH, Long Q, Apweiler R et Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 25:2865-2871.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marçais G, Roberts M, Subramanian P, Yorke JA, et Salzberg SL. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology.10:R42.

Autres publications de l'auteur

Annexe 1 :

Rapid Discovery of *De Novo* Deleterious Mutations in Cattle Using Genome Sequence Data: Enhancing the Value of Farm Animals as Model Species

E. Bourneuf[1,2,*], P. Otz[3,*], H. Pausch[4], V. Jagannathan[5], **P. Michot**[1,6], C. Grohs[1], G. Piton[1,2], S. Ammermüller[4], M.-C. Deloche[1,6], S. Fritz[1,6], H Leclerc[1,7], C. Péchoux[1,8], A. Boukadiri[1], R. Saintilan[1,6], F. Créchet[1,2], M Mosca[9], D. Segelke[10], F. Guillaume[1], S. Bouet[1], A. Baur[1,6], A. Vasilescu[11], L. Genestout[11], A. Thomas[12], A. Allais-Bonnet[1,6], D. Rocha[1], M.-A. Colle[13,14], C. Klopp[15], D. Esquerré[16,17], C. Wurmser[4], K. Flisikowski[18], H. Schwarzenbacher[19], J. Burgstaller[20], M. Brügmann[21], E. Dietschi[5], N. Huth[22], M. Freick[23] S. Barbey[24], G. Fayolle[25], C. Danchin-Burge[5], L. Schibler[5], B. Bed'Hom[1], B.J. Hayes[26,27], H. D. Daetwyler[26,27], R. Fries[4], D. Boichard[1], D. Pin[10], C. Drögemüller[5], A. Capitan[1,5]

ABSTRACT:
In humans, the clinical and molecular characterizations of sporadic syndromes are often hindered by the small number of patients and the difficulty to develop animal models for severe dominant conditions. Here we show that in livestock the availability of large data sets of whole-genome sequences, high-density SNP chips genotypes and phenotypic records offers unprecedented opportunity to dissect in record time the genetic architecture of phenotypes. We report the identification of eight dominant *de novo* mutations in CHD7, COL1A1, COL2A1, COPA, MITF, and REEP6 and take advantage of the structure of cattle populations to describe their clinical consequences and map modifier loci. In addition, we demonstrate the feasibility of anticipating the emergence of recessive genetic defects by detecting *de novo* deleterious mutations in the genomes of artificial insemination bulls. These results increase the attractiveness of cattle to confirm the genetic etiology of isolated clinical case reports and become a model species in the post genomic era.

Annexe 2 : Liste des articles de synthèse dans le numéro spécial de INRA Productions Animales

Duchesne A, Grohs C, **Michot P**, Boichard D, Floriot S, Fritz S., Capitan A. 2016. Du phénotype à la mutation causale d'anomalies récessives. INRA Productions Animales, 29, 319-328.

Fritz S., **Michot P**, Hoze C., Grohs C, Barbat A., Boussaha M., Boichard D, Capitan A. 2016. Anticiper l'émergence d'anomalies génétiques grâce aux données génomiques. INRA Productions Animales, 29, 339-350.

Boichard D, Grohs C, **Michot P**, Danchin C, Capitan A, Genestout L, Barbier S., Fritz S. 2016. Prise en compte des anomalies génétiques en sélection. INRA Productions Animales, 29, 351-358.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)
AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Titre : Utilisation des séquences de génome complet pour l'identification de mutations délétères responsables d'anomalies génétiques récessives chez le bovin

Mots clés : : génomique, mutations délétères, anomalies, bovins

Résumé :

L'effectif génétique réduit des races bovines entraîne une augmentation de consanguinité de l'ordre de 1% par génération et une forte dérive génétique. Cette évolution favorise l'émergence régulière d'anomalies génétiques récessives dans les populations, qu'elles soient de races laitières ou allaitantes. En France, l'Observatoire National des Anomalies Bovines (ONAB) a été créé dans le but de détecter et contrôler ces anomalies émergentes. Cependant, la détection par les observatoires sous-entend d'une part une diffusion large de l'allèle défavorable dans la population et d'autre part que l'anomalie présente un phénotype avec un tableau clinique spécifique permettant une déclaration du phénotype à l'ONAB. L'impact des anomalies génétiques est donc encore largement sous-estimé. Toutefois, le développement des technologies de génotypage et de séquençage de génomes, associés à l'ensemble des informations disponibles permet une détection efficace des mutations causales. Ainsi, l'objectif de cette thèse a été d'utiliser l'ensemble des données disponibles (phénotypes, génotypes, séquences, annotations fonctionnelles…) pour identifier et valider des mutations délétères ségrégant dans races bovines laitières et allaitantes françaises. Nous avons exploré différentes stratégies classiques - cartographies par homozygotie, haplotypes en déficit en homozygotes - qui gagnent en efficacité grâce aux données de séquence de génome complet (WGS). Nous avons également mis en place des approches alternatives de génétique inverse, basées sur l'exploitation des données WGS française et du consortium 1000 bull genomes.

Les travaux réalisé ont permis d'identifier les mutations causales associées à deux syndromes récessifs rapportés à l'ONAB : l'épidermolyse bulleuse jonctionnelle en race Charolaise (ITGB4, g.chr19: g.56488278_56493087del) et d'épilepsie idiopathique (MTCL1, g.chr24:41661691 G>A) récemment émergée dans la race Parthenaise. Nous avons également démontré la forte association entre l'haplotype MH1 et un polymorphisme affectant le gène PFAS (Chr19:g.28511199C>T ; p.R1205C). Par les stratégies de génétique inverse, nous avons également identifié une mutation probablement responsable de mortalité embryonnaire en race Normande affectant le gène CAD (Chr11:g72399397, p.Y452C) ainsi qu'une mutation affectant le gène RP1 (Chr14:g.23995411_23995412insA, p. R791KfsX13) responsable d'une dégénérescence progressive de la rétine et qui ségrége à forte fréquence en race Normande mais aussi dans d'autres races bovines européennes. Ces études encore en cours, fournissent également un inventaire des variants génétiques potentiellement délétères dont la caractérisation de l'effet sur le phénotype pourra être explorée. Enfin, l'identification de ces anomalies et des mutations délétères responsables ont abouti à la mise à disposition de tests de diagnostic efficaces pour permettre une contre-sélection raisonnée de ces variants délétères dans les populations bovines.

Title : Use of whole genome sequence data to identify new deleterious mutations associated to genetic defects in French dairy and beef cattle breeds.

Keywords: genomic, deleterious mutations, anomalies, bovine

Abstract

The reduced genetic size of cattle breeds leads to an increase in inbreeding of nearly 1% per generation and a strong genetic drift. This evolution favors regular emergence of recessive genetic abnormalities in dairy and beef cattle breeds. In France, the National Observatory of Bovine Anomalies (ONAB) was created with the aim to detect and control these new emergences. However, detection by the observatories implies a broad diffusion of the deleterious allele in the population and a phenotype with specific clinical features, allowing reporting to ONAB. Therefore the impact of genetic abnormalities is still largely underestimated. However, development of genotyping and genome sequencing technologies, coupled with all available information, makes possible effective detection of causal mutations. Thus, the aim of this thesis was to use all the available data (phenotypes, genotypes, sequences and functional annotations) to identify and validate deleterious mutations segregating in French dairy and beef cattle breeds. We used homozygosity mapping, and study of haplotypes with deficit in homozygous, two strategies boosted by using whole genome sequence (WGS) data. We also explored alternative reverse genetics strategies, based on data mining of French and 1000 bull genomes consortium WGS data.

In these different studies, we identified the causal mutations associated with two described recessive syndromes: epidermolysis bullosa junctional in Charolais race (ITGB4, chr19: g.56488278_56493087del) and idiopathic epilepsy (MTCL1, chr24:g.41661691G>A) recently emerged in the Parthenaise breed. We also demonstrated a strong association between the embryonic lethal mutation MH1 and a polymorphism affecting the *PFAS* gene (chr19: g.28511199C> T ; p.R1205C) in Montbeliarde cattle. With reverse genetic strategies, we identified another mutation, which affects the CAD gene (chr11: g72399397, p.Y452C), likely responsible for embryonic mortality in Normande cattle, and a frameshift mutation in *RP1* (chr14:g.23995411_23995412insA, p. R791KfsX13) responsible for progressive retinal degeneration segregating with a high frequency in the Normande breed but also in other European cattle breeds. These studies, still in progress, provide an inventory of potentially deleterious genetic variants, whose characterization could be explored in the future. At last, identification of mutations responsible for these genetic abnormalities provided or will provide diagnostic tests and allow efficient counter selection of these variants in French beef and dairy cattle populations.