Etude des polymorphismes altérant la régulation de l'expression des gènes chez le bovin

Gabriel Guillocheau

université PARIS-SACLAY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

NNT : 2018SACLA045

Etude des polymorphismes altérant la régulation de l'expression des gènes chez le bovin

Thèse de doctorat de l'Université Paris-Saclay préparée à AgroParisTech (Institut des Sciences et Industries du Vivant et de l'Environnement)

Ecole doctorale n°581 Agriculture, alimentation, biologie, environnement et santé (ABIES)

Spécialité de doctorat : Biochimie et biologie moléculaire

Thèse présentée et soutenue à Paris, le 19 décembre 2018, par

GABRIEL GUILLOCHEAU

Composition du Jury :

Alexandre Pery ICPEF, AgroParisTech	Président
Carole Charlier PhD, Université de Liège (GIGA CHU)	Rapporteure
Cédric Notredame PhD, Centre for Genomic Regulation (CRG)	Rapporteur
Gwenola Tosser-Klopp Directeur de recherche, INRA de Toulouse (GenPhySE)	Examinatrice
Dominique Rocha Directeur de recherche, INRA de Jouy-en-Josas (GABI)	Directeur de thèse

Thèse de doctorat

À André Guillocheau,

à Aliénor

Remerciements

Je ne sais pas s'il existe une bonne habitude pour la thèse ! Faut-il trimer jour et nuit shooté au café ou prendre le temps dans la thèse pour avoir du recul sur des sujets pas tout à fait banals ? À cette question, je n'ai pas de réponse à donner. Ma seule certitude c'est que seul je ne serais pas arrivé aussi loin. Il est donc normal de remercier ceux qui m'ont apporté un grand bonus dans ma thèse.

Tout d'abord, je tiens à remercier Claire Rogel-Gaillard, directrice de l'unité GABI (Génétique Animale et Biologie Intégrative) et Didier Boichard animateur de l'équipe G2B (Génétique et Génomique Bovine) pour m'avoir accepté au sein de cette équipe.

Je remercie bien évidemment mon directeur de thèse Dominique Rocha qui a su m'accompagner tout au long de ma thèse. Merci Dominique pour toutes les opportunités que tu m'as offertes (congrès, mission à l'étranger, collaborations), les conseils que tu m'as prodigués et ta grande exigence qui m'a permis de forger mon caractère et d'acquérir en rigueur. Merci surtout pour ton aide pour le deuxième article qui m'a appris à me débrouiller seul et à ne compter que sur moi.

Je tiens à remercier le jury d'ABIES qui m'a donné sa confiance sur la base de mes 10 minutes de présentation inattendues pour être éligible à une de leurs bourses et donc à ABIES et AgroParisTech pour le financement de ma thèse.

Je remercie chaleureusement les membres de mon jury pour l'évaluation de mon travail : tout d'abord Claire Charlier et Cédric Notredame qui en tant que rapporteurs auront la primeur de ces remerciements, Gwenola Tosser-Klopp et Alexandre Pery. Je souhaite aussi remercier les membres de mon comité de thèse pour leurs conseils et l'encadrement au cours de ma thèse : Jordi Estellé, Sandrine Lagarrigue, Andréa Rau et Daniel Zerbino.

Bien entendu, je remercie toute mes collègues de l'équipe G2B, les collègues de l'équipe PSGEN. Je remercie les membres du pôle RH : merci à Yvelise Fricot, Alexandra Vincent et Nathalie Lenoir pour votre aide dans tous les aspects administratifs (frais de missions, ordre de mission, *etc.*). Merci aussi à Fabien, Bruno et Thierry pour leur aide informatique et toutes ces discussions techniques sur l'informatique, je fus ravi d'avoir eu des interlocuteurs calés sur ces thématiques. Merci aussi Thierry d'avoir été mon partenaire de blague, même si ton absence a pesé sur la fin de ma thèse, je suis ravi que tu aies pu quitter la région parisienne.

Merci aux anciens doctorants de GABI qui m'ont apporté leurs expériences de la thèse et leurs bravos : Romain (le roi du scud), Alexis, Iola, Tatiana, Sonia et Jean-Noël. Je souhaite bonne chance à tous les futurs doctorants qui soutiendront bientôt et je leur transmet tous mes vœux de réussite : Rabia, Séb (team Sud-Ouest, chocolatine, rôse), Louise, Clémentine, Clémence, Marie, Anna-Charlotte, Audrey et Maria. Parmi mes collègues, je tiens à remercier toutes les autres personnes qui comme les précédentes m'ont apporté leur amitié : Julie, Maëlle, Rachel, Amandine, Chris, Eric et Sophie-Nancier (la personne la plus gentille et adorable que je connaisse). Clin d'œil à Jémane Bal'eikh qui se reconnaîtra !

Je remercie également les collègues qui m'ont accueillies pendant 3 mois en Ecosse au Roslin Institute et surtout Rachel, Lucas et Emily. Je remercie aussi David Hume, chef de mon équipe d'accueil au Roslin Institute pour ses conseils et son expertise.

Durant ma thèse, j'ai été sélectionné au label EIR-A (Ecole Internationale de Recherche - Agreenium) et j'y ai rencontré des gens formidables. Je tiens à remercier particulièrement toute la team *auriculus S.* qui s'est formé lors du séminaire à Montpellier : Juanito, Norman, Puikette, Nathan, Alice, Stephounette. Notre soutien mutuel tout au long de nos thèses est un des meilleurs moteurs que j'ai eu pour continuer, merci d'avoir été là pendant la phase la plus dure de mon existence et de ma thèse.

J'ai aussi participé durant ma thèse à l'association Doc'J dont je tiens à remercier tous les membres et surtout la team évènement. Merci Aline pour ton rire communicatif et ta persévérance. Merci à Elise d'avoir squatté mon mariage créant ainsi une amitié que j'espère éternelle. Merci Yuuki, pour ton flegme et ton calme mais aussi pour ta grande écoute vis à vis de l'overthinking personne que je suis. Je tiens aussi à remercier les habitués de nos événements : Déborah, Baudoin, Bastien et Yohann. Merci à Thé, pour ta douceur, ta gentillesse et ton humour. Je remercie le DJ Teddy B. pour avoir animé certaines de nos soirées (je ne dis pas qu'Alexis est Teddy B. mais je n'ai jamais vu les deux au même endroit au même moment).

Il m'est très difficile de faire des remerciements sans penser à ma famille. J'embrasse mes parents pour leur soutien dans mes études même s'ils ne comprenaient pas forcément tout ce que je faisais. J'embrasse aussi mes quatre sœurs qui à défaut d'être adorables sont extraordinaires, votre frère est fier de vous. J'embrasse les autres membres de ma famille. Gros câlin à ma petite Zelda partie trop tôt et à mon chat Wolfram. Enfin, toutes mes pensées vont à feu mon grand-père André Guillocheau à qui cette thèse est dédiée.

Je tiens à remercier tous les amis qui m'ont soutenu durant ces trois années. Merci à la team physique-chimie avec ma Crapine, Wiwi, mon Jeannot, Joris, Audrey et Jill. Merci à Coralie pour ton rire, ta créativité et ta folie. Merci à Swale pour ta labélisation de mes blagues et ton humour. Merci à mes deux témoins de mariage Samos et Keiji, je vous aime les gars.

Laure, merci pour ton soutien indéfectible même quand j'étais ronchon ou de mauvais poil. Merci pour ta grande écoute, pour ton épaule toujours tendue et pour ta grande affection. Je te souhaite d'être très heureuse. Continue de cultiver ta créativité, c'est ta plus grande force.

Roxane, ma super co-bureau, je suis heureux de t'avoir comme amie, sans ton soutien quotidien je n'aurais pas parcouru autant de chemins. Merci de m'avoir écouté, secoué, conseillé même quand j'étais extrêmement énervant. Merci de m'avoir supporté jusqu'au bout en recouvrant régulièrement mon tableau de petits cœurs. Merci pour tous ces instants de délires, de complicités et d'amitié. Désolé je ne serais plus là pour te nourrir /) /) /). Tes petits mots mignons, ce beau rite, vont me manquer !

Pour finir, je tiens à remercier ma merveilleuse femme. Merci Aurélie de m'aimer et de m'apporter soutien dans les meilleurs moments comme dans les pires et pas juste parce que c'est mentionné dans les articles de loi du mariage. Merci d'être mon meilleur public mais aussi d'arriver encore à me faire rire et à me surprendre par ta folie douce depuis toutes ces années. Je t'aime !

Voici une petite citation qui résume très bien ma thèse et pour les plus lettrés, des contrepets facétieux se sont glissés dans ces remerciements :

> *« La vie, c'est comme un escargot. On porte un lourd fardeau sur le dos, il faut en baver pour avancer, et ça laisse toujours des traces. »*
> Stéphane Heska

Table des matières

Remerciements ii

Table des matières v

Table des figures vii

Liste des tableaux vii

Liste des abréviations viii

Introduction 1

A Expression des gènes : les différentes étapes de la régulation 1

A.1 La transcription : de l'ADN à l'ARN . . . 1
A.2 Les modifications post-transcriptionnelles : variations et stabilisation de l'ARN . . . 3
A.3 Le transport de l'ARNm : du noyau au cytoplasme . . . 10
A.4 La traduction : la synthèse de la protéine . . . 11
A.5 Dégradation de l'ARNm . . . 15
A.6 Modification post-traductionnelle : activation de la protéine . . . 16
A.7 Les microARN : leurs rôles dans l'expression des gènes . . . 20
A.8 Les longs ARN non codants : des ARN très polyvalents . . . 23
A.9 Rôle de l'épigénétique dans l'expression des gènes . . . 24

B Étudier les mécanismes de régulation de l'expression des gènes 28

B.1 Séquençages des génomes et des transcriptomes . . . 28
B.2 Importance des polymorphismes : détection des variants . . . 34
B.3 Etude bio-informatique de la régulation de l'expression des gènes . . . 37
B.4 Validation expérimentales de la régulation de l'expression des gènes . . . 41

Stratégies de la thèse 48

Article 1 : Etude de l'expression allèle-spécifique dans le muscle bovin 50

Article 2 : Etude de l'expression allèle-spécifique intra individu et inter tissu chez la vache holstein. 86

A Introduction 86

B Matériels et méthodes 87

B.1 Animaux et échantillonnage . . . 87

B.2 Séquençage génome entier (WGS) et alignement de séquence . . . 87
B.3 RNA-Seq et alignement de séquence . . . 88
B.4 Identification et annotation des SNPs . . . 89
B.5 Détection des ASE-SNPs . . . 90
B.6 Analyse de l'expression tissu-spécifique . . . 90

C Résultats et discussions **90**
C.1 Statistiques sur les données de séquençage ARN et ADN . . . 90
C.2 Détection des variants . . . 91
C.3 Comparaison des SNPs . . . 93
C.4 Identification des ASE-SNPs . . . 93
C.5 Expression tissu-spécifique . . . 93

Discussion générale 97

A Apports des résultats **97**

B Retour sur les approches employées **97**
B.1 Validation à approfondir . . . 97
B.2 Approche ASE . . . 98
B.3 Tests et améliorations . . . 99

Bibliographie **103**

Annexe A : Rapport scientifique de la mission de 3 mois au Roslin Institute **119**

Publications, Communications et Encadrement 124

Table des figures

1 Complexe de préinitiation de la transcription de l'ARN polymérase II par Krishnamurthy et Hampsey (2009) . . . 2
2 Terminaison de la transcription chez l'Homme (Rosonina *et al.*, 2006) . . . 4
3 La réaction d'épissage de l'ARNm précurseur par McManus et Graveley (2008). . . . 5
4 Les différents mécanismes d'épissage alternatif inspiré par Keren *et al.* (2010). . . . 7
5 Modèle de l'exportation de l'ARNm par Cheng *et al.* (2006). . . . 11
6 Modèle canonique de la voie de l'initiation de la traduction chez les eucaryotes par Jackson *et al.* (2010) . . . 13
7 Phase d'élongation de la traduction chez les Eucaryotes par Fritz et Boris-Lawrie (2015). . . . 14
8 Voie canonique de la biogénèse des microARN par Jung et Suh (2015) . . . 21
9 Voies alternatives de la biogènese des microRNAS par Ha et Narry Kim (2014) . . . 23
10 Les différents niveaux de condensation de l'ADN par Annunziato (2008) . . . 25
11 Evolution du coût de séquençage en comparaison à loi de Moore par Hayden (2014). . . . 28
12 Résumé du séquençage haut-débit pour les 4 méthodes principales de NGS. . . . 30
13 Fonctionnement des différentes technologies TGS par Schadt *et al.* (2010). . . . 32
14 Protocole d'analyse pour la découverte des petits variants (SNPs et indels) de lignée germinale avec les outils de GATK d'après le best practices workflow. . . . 36
15 Conséquences prédites par VEP (McLaren *et al.*, 2016). . . . 37
16 Mécanismes des variations régulatrices en local (cis) et en distant trans par (Skelly *et al.*, 2009). 40
17 Aperçu de l'approche MPRA selon Melnikov *et al.* (2012). . . . 43
18 Base moléculaire de la réaction de pyroséquençage par Royo *et al.* (2007) . . . 44
19 Diagramme du protocole de pyroséquençage utilisant une amorce M13 universel biotinylée par Royo *et al.* (2007). . . . 45
20 Principe de l'analyse ChIP-Seq par Mundade *et al.* (2014). . . . 46
21 Technologies de Genome Editing exploitant la réparation endogène de l'ADN par Hsu *et al.* (2014). . . . 47
22 Stratégie globale de la thèse. . . . 49
23 Diagramme du pipeline pour détecter des ASE-SNPs. . . . 50
24 Clustering hiérarchique et heatmap des échantillons à partir des données d'expression des gènes. 95
25 Clustering hiérarchique et heatmap des échantillons à partir de la présence d'ASE-SNPs par gène. 96
26 Diagramme de l'expression allèle spécifique (ASE), de l'ASE lié à une influence parentale (PO-ASE) et du locus de caractères quantitatifs associé à l'expression (eQTL). . . . 101
27 Schéma de la détection des variants causaux en couplant les approches ASE et eQTL avec (A) la détection des TFBS (Transcription Factor Binding Site) et (B) la détection des mirBS (microRNA Binding Site). . . . 102
28 Workflow d'analyse du single-cell RNA sequencing. . . . 102

Liste des tableaux

1 Rôles des ARN polymérases. . . . 2
2 Table des codons ARN. . . . 15
3 Modifications post-traductionnelles fréquentes par acide aminé. . . . 19
4 Comparaison des différentes générations de séquençage traduit de Schadt *et al.* (2010). . . . 33
5 Résultats de l'alignement des lectures WGS . . . 91
6 Résumé des annotations VEP des SNPs détectés à partir des données ADN (WGS) et ARN (RNA-Seq). . . . 92
7 Distribution des ASE-SNPs détectés chez les vaches holstein. . . . 94

Liste des abréviations

ABCE1 : ATP-binding cassette sub-family E member 1
ADAR : Adenosine Deaminase Acting on RNA
ADN : Acide désoxyribonucléique
AID : Activation-Induced Cytidine deaminase
APOBEC : Apolipoprotein B mRNA editing Enzyme, Catalytic polypeptide-like
ARN : Acide ribonucléique
ARNm : ARN messager
ARN pol II : ARN polymerase II
ARNr : ARN ribosomique
ARNt : ARN de transfert
BCF : Variant Calling Format
Cas9 : CRISPR associated protein 9
CDC : Charge-Coupled Device
ChIP : ImmunoPrécipitation de Chromatine
CNV : Copy Number Variation
CoTC : CoTranscriptional Cleavage
CPB : Cap-Binding Protein
CpG : Cytosine-phosphate-Guanine
CRISPR : Clustered Regularly Interspaced Short Palindromic Repeats
CTD : Carboxyl Terminal Domain
DGCR-8 : DiGeorge syndrome Critical region Gene-8
DNMT : ADN méthyl-transférases
DSB : DNA Double-Strand Breaks
dSNP : driver SNP
eEF : eukaryotic Elongation Factor
eGWAS : expression Genome-Wide Association Studies
eIF : eukaryotic Initiation Factor
EJC : Exon Junction Complex
eQTL : expression Quantitative Trait Locus
eRF : eukaryotic Release Factor
ESE : Exonic Splicing Enhancer
ESS : Exonic Splicing Silencer
EXP5 : Exportine 5
G2B : Génétique et Génomique Bovine
GABI : Génétique Animale et Biologie Intégrative
GFP : Green Fluorescent Protein
Gln : Glutamine
GDP : Guanosine Diphosphate
DGCR-8 : DiGeorge syndrome Critical region Gene-8
GMP : Guanosine Monophosphate
GTP : Guanosine Triphosphate
GVF : Genome Variation Format
HAT : Histone AcétylTransférase
HDAC : Histones Désacétylases
HDR : Homology-Direct Repair
HMT : Histone MéthylTransférase
IRES : Internal Ribosome Entry Site
ISE : Intronic Splicing Enhancer
ISS : Intronic Splicing Silencer
LD : Linkage Desequilibrium
lincRNA : long intergenic non coding RNA

lncRNA : long ARN non-codant (long non coding RNA)
MAPK : Mitogen-Activated Protein Kinase
miARN : microARN
mirBS : microRNA Binding Site
MPRA : Massively Parallel Reporter Assay
mRNP : Messenger RiboNucleoProtein
NATs : N-terminal acétyltransférases
NGS : Next Generation Sequencing
NHEJ : Non-Homologous End-Joining
PABP : Poly-A Binding Protein
PAM : Peptidylglycine alpha-Amidating Monooxygenase
pARNi : petits ARN interférents
PCR : Polymerase Chain Reaction
Pré-miARN : microARN précurseur
Pri-miARN : microARN primaire
PRMTs : Protein Arginine MethylTransférases
RBP : RNA Binding Protein
RH : Recombinaison Homologue
RISC : RNA-induced silencing complex
rSNP : SNP régulateur
snoARN : petit ARN nucléolaire (Small nucleolar RNAs)
snRNP : petit ribonucléoprotéine nucléaire (small nuclear ribonucleoprotein)
SOLiD : Sequencing by Oligo Ligation Detection
STARR-seq : Self-Transcribing Active Regulatory Region sequencing
SUMO : Small Ubiquitin MOdifier of proteins
TALE : Transcription Activator-Like Effectors
TALEN : Transcription Activator-Like Effectors Nuclease
TAR : TransActivation Region
TBP : protéine de fixation à la boîte TATA (TATA box-binding protein)
TF : Facteur de transcription (Transcription Factor)
TFBS : Transcription Factor Binding Site
TFII : Facteurs de transcription de l'ARN polymérase II
TRBP : TAR RNA-Binding Protein
TREX : Transcription Export complex
tSNP : transcript SNP
TUTase : Terminal Uridylyl Transferases
UTR : Untranslated Transcribed Region
VCF : Variant Calling Format
WGS : Whole Genome Shotgun
Xist : X inhibitory specific transcript
ZF : Zinc Finger
ZFN : Zinc Finger Nuclease

Introduction

A. Expression des gènes : les différentes étapes de la régulation

L'expression des gènes désigne l'ensemble des processus biochimiques conduisant à la production d'ARN (acide ribonucléique) codant ou non codant et de protéines à partir de l'information stockée dans les gènes. De la séquence d'ADN (acide désoxyribonucléique) à la protéine produite, de nombreux mécanismes sont mis en place pour réguler l'expression des gènes afin de permettre la différenciation cellulaire, l'expression tissu-spécifique ou l'adaptabilité de l'organisme à son environnement. Durant ce chapitre, nous allons détailler les étapes-clés de la régulation : de la transcription à la traduction en passant par le contrôle des ARN non codants et le rôle de l'épigénétique.

A.1. La transcription : de l'ADN à l'ARN

La transcription est la synthèse de l'ARN à partir d'une séquence d'ADN présent dans un gène. Chez les eucaryotes, la transcription se déroule dans le noyau en 3 étapes : l'initiation, l'élongation et la terminaison.

A.1.1. L'initiation

L'initiation est la phase de fixation de l'ARN polymérase à l'ADN. Il existe 5 polymérases différentes et c'est l'ARN polymérase II qui est impliquée pour la synthèse des précurseurs d'ARNm (Table 1). Dans le cas de l'ARN polymérase II (ARN pol II) pour la synthèse de l'ARNm, cette fixation s'effectue au niveau du promoteur par le biais de nombreux cofacteurs protéiques qui forment le complexe d'initiation. Ce complexe est formé de la TBP (TATA box-Binding Protein) qui reconnaît la boîte TATA du promoteur du gène (Geiger *et al.*, 1996) et de facteurs de transcription de l'ARN polymérase II (les TFII - Orphanides *et al.*, 1996). Ces facteurs de transcription vont permettre l'initiation de la transcription, notamment en permettant l'ouverture de la double hélice (TFIIH - Drapkin *et al.*, 1994) et permettre la reconnaissance précise du site d'initiation (TFIIB - Pinto *et al.*, 1992). Ce complexe d'initiation est suffisant pour déclencher une activité transcriptionnelle mais elle est faible (Weil *et al.*, 1979). D'autres facteurs spécifiques vont agir sur le complexe d'initiation pour influencer cette activité basale en l'amplifiant ou en l'inhibant. Ces protéines activatrices ou inhibitrices vont se fixer à des promoteurs distaux spécifiques (séquences *cis*-régulatrices) de l'ADN

(Figure 1). Ces promoteurs sont appelés silenceurs (silencers) quand ils recrutent des cofacteurs inhibiteurs et amplificateurs (enhancers) quand ils recrutent des cofacteurs activateurs. Ces promoteurs distaux peuvent se situer à des milliers de nucléotides du promoteur proximal et ils vont pouvoir agir grâce à la courbure de l'ADN (Roberts *et al.*, 1993). Dans le cas des enhancers, une fois liées les protéines activatrices peuvent réguler la synthèse de l'ARNm : soit en induisant un changement de conformation du complexe d'initiation par le biais d'un médiateur, soit en éliminant les répresseurs du promoteur, soit en facilitant le recrutement des éléments du complexe d'initiation (facteurs généraux et l'ARN pol II) sur le promoteur, soit en favorisant l'échappée de l'ARN pol II permettant le démarrage de la phase d'élongation (Roberts *et al.*, 1993).

ARN polymérase	ARN synthétisé	Référence
ARN polymérase I	ARN précurseur 45S maturant en ARNr 28S, 18S et 5,8S	Grummt (1998)
ARN polymérase II	ARNm précurseurs et la plupart des snoARN et des microARN	Lee *et al.* (2004)
ARN polymérase III	ARNt, ARNr 5S et les autres petits ARN trouvés dans le noyau et le cytosol	Willis (1993)
ARN polymérase IV	pARNi (spécifique aux plantes)	Herr *et al.* (2005)
ARN polymérase V	ARN impliqués dans la méthylation de l'ADN en guidant les pARNi	Wierzbicki *et al.* (2009)

TABLE 1 – Rôles des ARN polymérases.

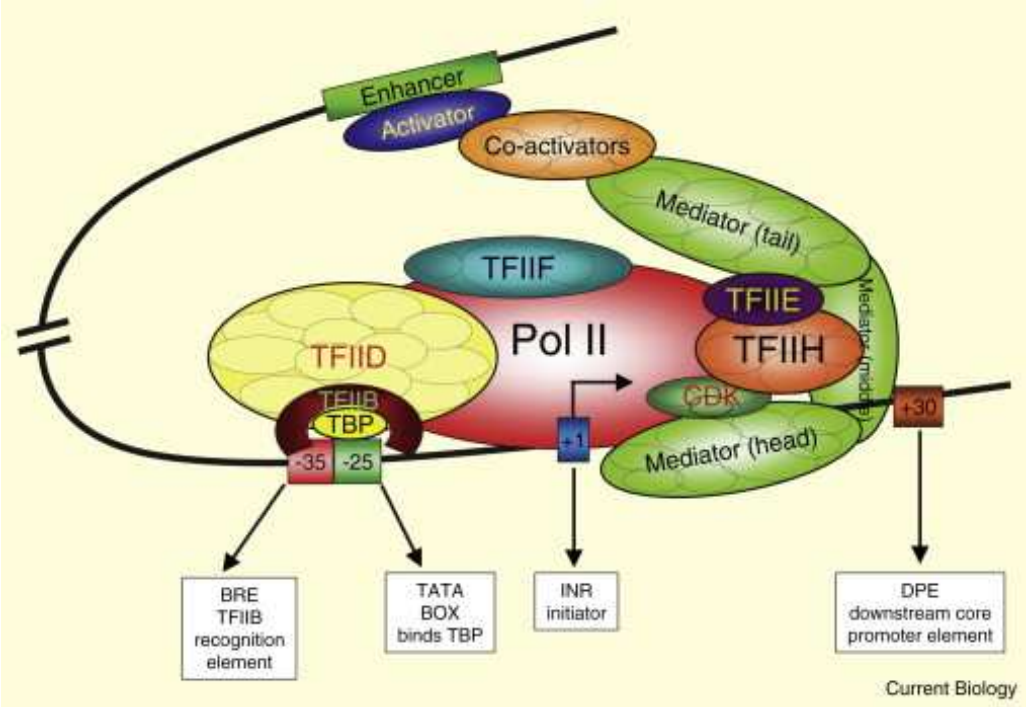

FIGURE 1 – Complexe de préinitiation de la transcription de l'ARN polymérase II par Krishnamurthy et Hampsey (2009)

A.1.2. L'élongation

L'élongation est la phase de lecture de l'ADN par l'ARN polymérase II produisant au fur et à mesure l'ARN précurseur. La phosphorylation du domaine CTD (Carboxy Terminal Domain – un domaine spécifique d'une sous-unité de l'ARN polymérase II) va déplacer l'ARN polymérase jusqu'au lieu d'origine de la transcription et démarrer l'élongation (Spencer et Groudine, 1990). Un ARN précurseur complémentaire du brin matrice de l'ADN (brin antisens) commence à être synthétisé selon la direction 5'-3' au rythme d'environ 25 nucléotides par seconde (Wickiser *et al.*, 2005). Les facteurs protéiques d'élongation de l'ARN pol II (TFIIF et TFIIS principalement) vont relâcher la structure des chromatines permettant ainsi la progression le long de l'ADN et l'écartement progressif de ses deux brins.

A.1.3. La terminaison

La terminaison : durant cette phase l'ARN polymérase II va être stoppée pour libérer l'ARN précurseur. Chez les eucaryotes, cette terminaison se déroule en parallèle de la polyadénylation en 3' de l'ARNm au niveau du signal de polyadénylation (séquence AAUAAA sur l'ARN synthétisé - Teixeira *et al.*, 2004).

Il existe deux modèles pour expliquer le processus de terminaison de l'ARN pol II : le modèle allostérique/anti-terminateur (Figure 2A) et le modèle Torpedo (Figure 2B). Dans le modèle allostérique (ou anti-terminateur), la terminaison est causée par la déstabilisation et/ou par les changements de conformation du complexe d'élongation de l'ARN pol II après la transcription du signal de polyadénylation. La dissociation du complexe de transcription est enclenchée soit par la libération de facteur d'anti-terminaison (Figure 2A, à gauche) soit par le recrutement d'un facteur de terminaison (Figure 2A, à droite). Dans le modèle Torpedo (Tollervey, 2004), le clivage au site de polyadénylation crée un site d'entrée pour une 5' -> 3' exonucléase (Xrn2 chez l'homme, West *et al.*, 2004) : le site CoTC (CoTranscriptional Cleavage). Xrn2 va ensuite dégrader l'ARN en aval du site de clivage. On obtient un ARNm précurseur qui va alors subir des étapes de maturation.

A.2. Les modifications post-transcriptionnelles : variations et stabilisation de l'ARN

L'ARN précurseur va être maturé en ARN messager avant d'être exporté hors du noyau. Cette phase de maturation de l'ARN se déroule en 3 étapes : l'addition de la coiffe, l'excision-épissage et l'ajout de la queue poly-A. En plus de cette maturation, l'ARN va aussi être adapté en fonction des

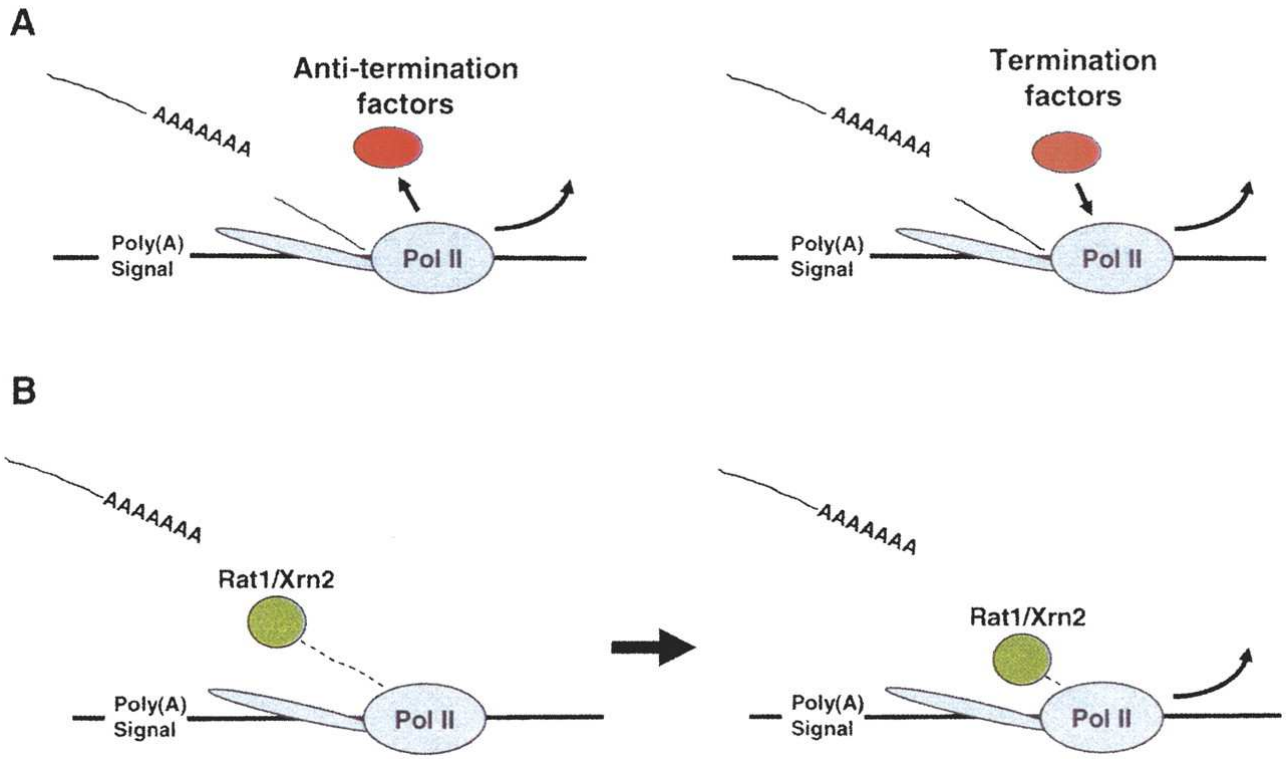

FIGURE 2 – Terminaison de la transcription chez l'Homme (Rosonina *et al.*, 2006)

besoins de l'organisme, soit de manière permanente avec l'épissage alternatif ou l'édition de l'ARN, soit de manière réversible avec la modification épigénétique de l'ARN (épitranscriptomique).

A.2.1. L'addition de la coiffe

Elle a lieu au début de la transcription alors que l'ARN précurseur ne compte pas plus de 30 nucléotides. Elle consiste en l'ajout d'une coiffe protectrice à l'extrémité 5' de l'ARNm afin de limiter la réactivité de cette extrémité (jonction non voulue par exemple) et sa reconnaissance par les exonucléases (protection contre la dégradation). Cette coiffe est formée par l'ajout d'une GMP (Guanosine Monophosphate) avec une méthylation sur l'azote 7 de la base (7-méthylguanosine ou m^7G) et par la méthylation en 2' du ribose du premier nucléotide et parfois du deuxième. Cette coiffe joue aussi un rôle dans l'exportation de l'ARNm vers le cytoplasme et lors de l'étape de l'initiation de la traduction (Bird *et al.*, 2016).

A.2.2. L'excision-épissage

Chez les eucaryotes, les gènes sont constitués d'exons (parties codantes du gènes) séparés par des introns qui sont intégralement recopiés dans l'ARN précurseur. Cet ARN va subir une phase d'excision des introns suivie d'une phase d'épissage (ligature bout à bout des exons) afin de former

l'ARNm. Ces phénomènes se déroulent au cours de la transcription. Les introns sont bornés par un site donneur à l'extrémité 5' (séquence consensus 5'-GU) et un site accepteur à l'extrémité 3' (séquence consensus 3'-AG). Dans la majorité des cas, les introns vont être épissés par le splicéosome mais certains vont s'auto-épissés.

Pour le premier cas, l'excision des introns est effectuée par une formation « en lasso » provoquée par le splicéosome (ovales oranges, Figure 3) composé de snRNP [1] combinés à d'autres protéines (Jurica et Moore, 2003). L'épissage est réalisé par réaction d'un nucléotide à adénine situé dans la boîte de branchement (une séquence interne de l'intron situé à une quarantaine de bases du site receveur) avec un nucléotide à guanine situé dans le site donneur. En découle un clivage entre l'intron et l'exon situé en amont et la formation d'une boucle d'ARN de la forme d'un lasso. Ensuite, le site accepteur en 3' est reconnu par deux snRNP du splicéosome et une liaison est créée avec l'extrémité 3' de l'exon en amont. L'épissage des deux exons libère l'intron excisé dont la structure sera ensuite ouverte pour être dégradé par des ribonucléases. Le splicéosome est ensuite décroché par la translocation en 3' vers 5' de la protéine PRP22 (McManus et Graveley, 2008).

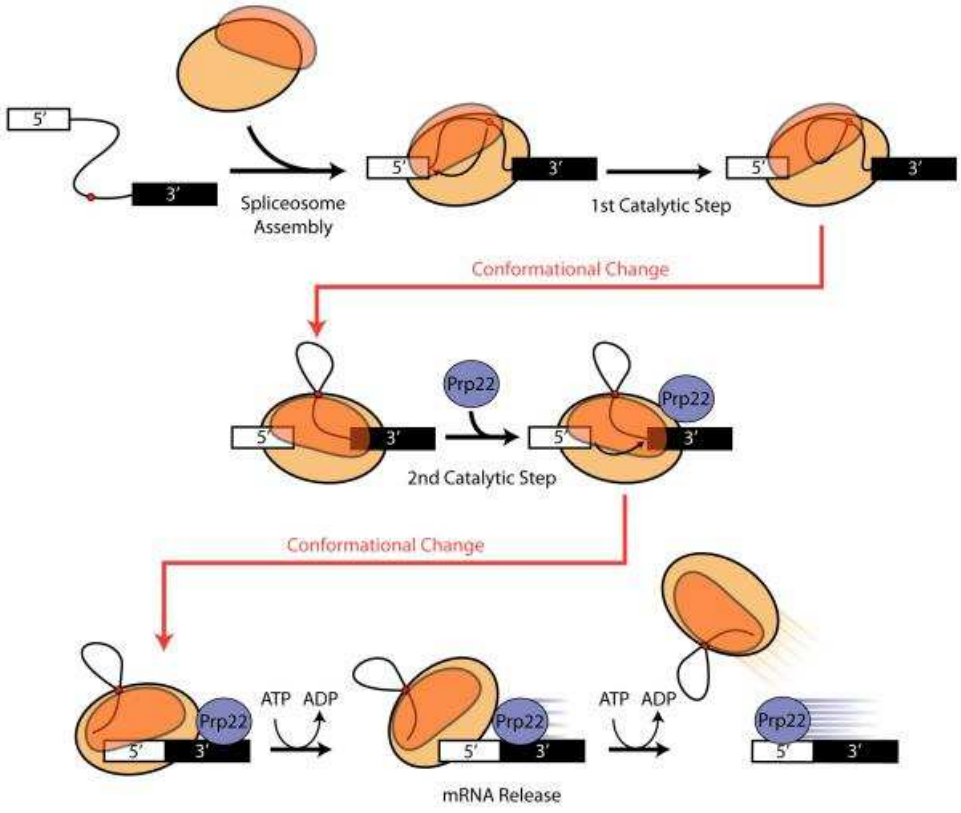

FIGURE 3 – La réaction d'épissage de l'ARNm précurseur par McManus et Graveley (2008).

Dans le cas des introns auto-épissables, il s'agît d'introns particuliers fortement structurés capables de s'épisser de manière autonome, c'est-à-dire sans l'aide du splicéosome ou d'autres com-

1. Les petites ribonucléoprotéines nucléaires (ou snRNP) sont des complexes mixtes entre des petits ARN nucléaires et des protéines.

plexes en *trans*. Ils sont regroupés en deux classes principales : les introns de groupe I (Vicens *et al.*, 2008) et les introns de groupe II (Pyle, 2016). Les introns de groupe II sont excisés sous forme de lasso comme lors de l'action du splicéosome alors que les introns de groupe I sont dégradés par l'intervention d'un cofacteur nucléotidique externe (GMP, GDP ou GTP).

A.2.3. L'épissage alternatif

L'épissage ne se produit pas toujours de manière constitutive, c'est-à-dire que l'ARNm créé n'est pas obligatoirement la succession de tous les exons présents. On parle d'épissage alternatif : l'excision des introns (ou des portions d'exons) peut permettre une suture d'exons différents. Ce mécanisme joue un rôle important dans la diversité de l'expression du gène en démultipliant ses capacités codantes (Keren *et al.*, 2010). L'épissage alternatif est régulé par des protéines de liaison à l'ARN (activateurs et répresseurs) se liant à des séquences régulatrices (enhancers ou silencers) et/ou par la structure secondaire de l'ARNm. Ces protéines régulatrices agissent en facilitant le recrutement du spliceosome ou en le recrutant. Les éléments *cis*-régulateurs présent dans l'exon affectent l'épissage de celui-ci (les Exonic Splicing Enhancer et les Exonic Splicing Silencer : ESE et ESS) et ceux présents dans l'intron interagissent sur les exons adjacents (les Intronic Splicing Enhancer et les Intronic Splicing Silencer : ISE et ISS) (Baralle et Baralle, 2018). Il existe différentes formes de cet épissage alternatif résumées dans la Figure 4 et un ARNm peut subir plusieurs formes d'épissage alternatif.

L'exon cassette

L'exon cassette est le mode le plus fréquent d'épissage alternatif chez les mammifères (Sammeth *et al.*, 2008). Un exon peut être épissé de l'ARN précurseur ou ne pas l'être. L'inclusion ou l'exclusion d'un exon permet une grande combinaison de transcrits. Par exemple, le gène *slo* code une protéine qui forme un canal au potassium activé par Ca^{2+} dans la cochlée des mammifères (Graveley, 2001). L'action de ce canal joue un rôle dans la perception des sons qui font vibrer les cils présents dans la cochlée. Il existe un exon alternatif dans l'ARN précurseur nommé STREX. Quand la protéine a été codée avec un ARNm contenant STREX elle a une plus grande sensibilité aux ions Ca^{2+} et une activité plus longue que la protéine codée sans.

Les exons mutuellement exclusifs

Un seul des deux exons est conservé lors de l'épissage. De nombreux mécanismes différents peuvent expliquer cette exclusion mutuelle (Pohl *et al.*, 2013). Par exemple, le transcrit précurseur

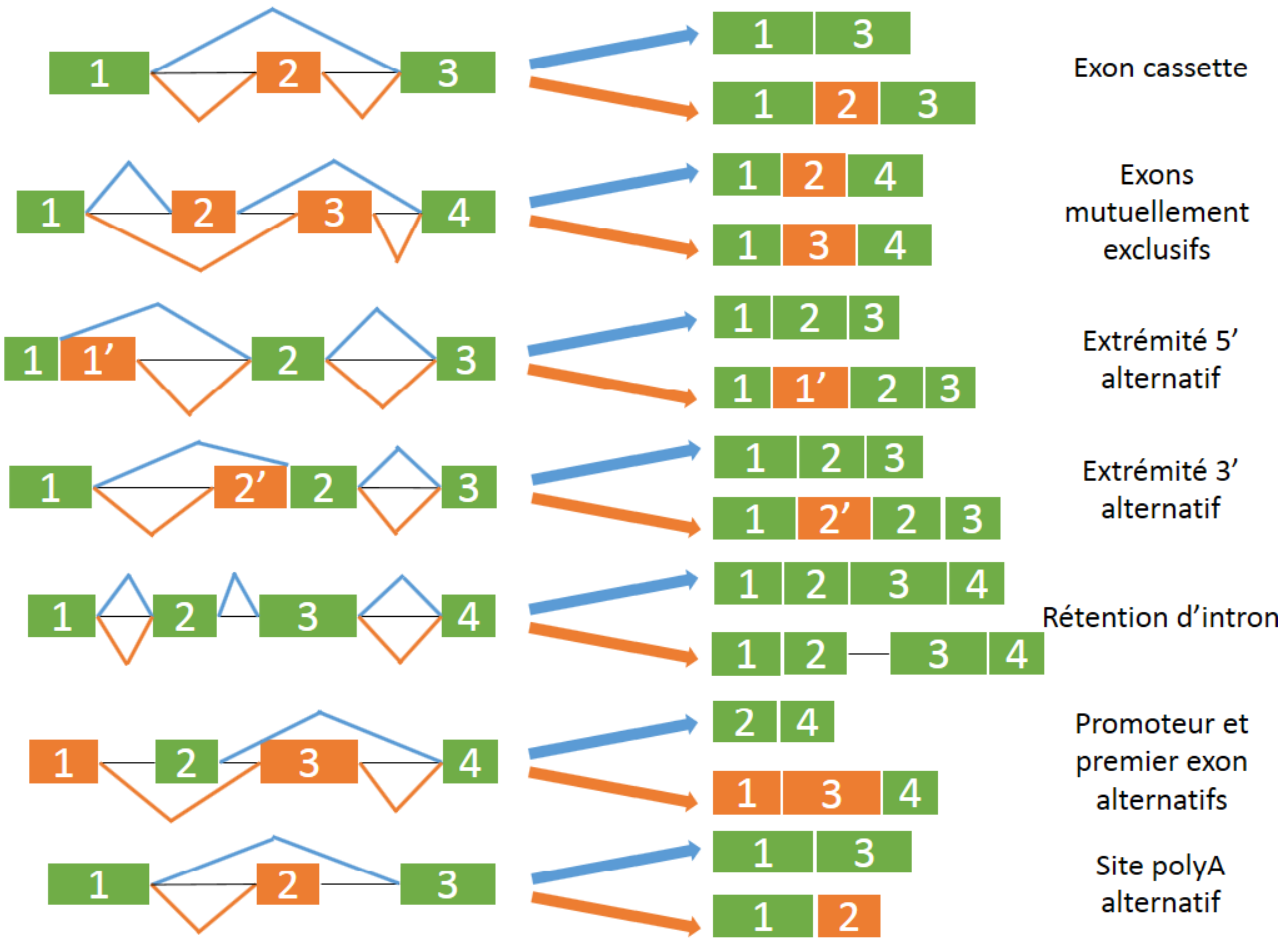

FIGURE 4 – Les différents mécanismes d'épissage alternatif inspiré par Keren *et al.* (2010).

de l' α-tropomyosine peut être traduit en 2 protéines isoformes possédant soit l'exon 2 soit l'exon 3. La protéine possédant l'exon 3 étant la plus fréquente dans les cellules, l'autre n'étant présente que dans les muscles lisses. Cette exclusion mutuelle est due à l'insertion ou non d'un élément d'espacement en 5' de l'intron entre l'exon 2 et l'exon 3 (Smith et Nadal-Ginard, 1989).

Site d'épissage alternatif en 5'

Il existe un site donneur alternatif changeant la fixation en 3' de l'exon en amont (Dou *et al.*, 2006). Par exemple, le gène *RAD9A* qui code une protéine de contrôle du cycle cellulaire possède un transcrit précurseur avec deux sites d'épissages alternatifs en 5'. Ces sites conduisent à une différence d'expression des deux transcrits (Bortfeldt *et al.*, 2008).

Site d'épissage alternatif en 3'

Il existe un site récepteur alternatif changeant la fixation en 5' de l'exon en aval (Dou *et al.*, 2006). Par exemple, le gène *Slack* code une protéine formant un canal ionique en s'associant avec

la protéine Slick. L'association de ces deux protéines dépend du domaine N-terminal de Slack qui possède deux transcrits avec une extrémité 3' alternative : *Slack-A* et *Slack-B* (Chen *et al.*, 2009). La protéine transcrite à partir de Slack-A ne se fixera pas à la protéine Slick contrairement à celle transcrite à partir de Slack-B.

La rétention d'intron

La séquence intronique peut être épissée en tant qu'intron ou être conservée dans l'ARNm final (Sakabe et de Souza, 2007). La présence de sites d'épissage non optimaux conduit généralement à cette rétention de l'intron. C'est le cas par exemple chez l'homme avec le gène β-globine 2 dont le transcrit précurseur peut conserver l'intron 2 (McCullough et Berget, 1997).

L'existence de promoteurs alternatifs

La position d'autres promoteurs peut changer le premier exon. Il s'agit plus d'une régulation transcriptionnelle que de l'épissage alternatif *stricto sensu* mais il en découle un ARNm différent du gène d'origine. De plus, la plupart des promoteurs influence l'épissage alternatif (Xin *et al.*, 2008).

L'existence d'un site poly-A alternatif

L'existence d'un site terminal alternatif peut changer l'exon terminal. Comme pour le mécanisme précédent, il s'agit en fait d'une régulation transcriptionnelle.

A.2.4. L'ajout de la queue poly-A

Le pré ARNm est d'abord clivé dans le noyau par un complexe protéique, au niveau d'un site consensus de polyadénylation (AAUAAA). Ce clivage est suivi d'une réaction de polyadénylation qui consiste en l'addition d'environ 200 résidus adénosine sur le produit de clivage en amont de la coupure, alors que le fragment en aval est rapidement dégradé. Cette queue poly-A confère de la stabilité au futur ARNm et se perd au fur et à mesure qu'il est traduit.

A.2.5. L'édition des ARN

En plus de l'épissage alternatif, il existe un autre mécanisme post-transcriptionnel qui apporte de la diversité dans l'expression des gènes : l'édition des ARN. Ce phénomène apparaît plus fréquemment dans l'ARNm que dans l'ARNr ou l'ARNt et il peut aussi se produire dans les microARN (Li et Mason, 2014). Il s'agit d'un processus radical qui modifie un nucléotide par un autre dans un codon pouvant entraîner le changement de l'acide aminé. Ce phénomène a été détecté chez les kinétoplastides

(des protistes parasites), dans des mitochondries et chloroplastes de plantes et dans des noyaux de mammifères (Gott et Emeson, 2000). Chez les mammifères, 2 mécanismes ont été recensés.

L'édition de l'adénosine en inosine (A-vers-I)

L'inosine est formée par désamination de l'adénosine. L'édition A-vers-I est catalysée par les enzymes ADAR (Adenosine Deaminase Acting on RNA - Nishikura, 2010) qui agissent sur l'ARN (ADAR1, ADAR2 et ADAR3). Dans la machinerie cellulaire et lors du séquençage, l'inosine qui n'est pas une base classique sera reconnue comme une guanine. Chez l'homme, seule une petite portion de ces sites d'édition se retrouve dans la séquence codant une protéine (2 411 sites sur les 1 379 403 sites annotés dans la base de données RADAR, Ramaswami et Li, 2014). Cette modification peut entraîner un changement du codon (Zinshteyn et Nishikura, 2009). Quand cette édition se déroule dans les régions non codantes, elle peut altérer les sites cibles reconnus par les microARN ou influencer l'épissage alternatif (Li *et al.*, 2009).

L'édition de la cytidine en uridine (C-vers-U)

L'uridine est formée par désamination de la cytidine. Cette édition est catalysée par la famille d'enzymes AID/APOBEC (Activation-induced cytidine deaminase et Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) qui compte AID, APOBEC1, APOBEC2 et APOBEC3.

Des recherches sur les mammifères ont montré que cette édition pouvait entraîner la transformation d'un codon de la glutamine (CAA) en un codon stop (UAA) (Fossat et Tam, 2014). L'exemple le plus courant (qui donne son nom à la famille APOBEC) est le couple ApoB100 et ApoB48 (Chen *et al.*, 1987; Amengual *et al.*, 2018). Il s'agit de deux apolipoprotéines (protéines qui assurent la cohésion des lipoprotéines, des protéines transportant les lipides dans l'organisme) codées par le même transcrit *APOB-201*. ApoB100 est une protéine plutôt lourde (500kDa avec 4 563 acides aminés) qui transporte les lipides dans le sang à la surface des lipoprotéines VLDL, IDL et LDL. Pour ApoB48, le transcrit a été édité sur le Gln2153 (codon 2153 codant la glutamine) créant ainsi un codon stop. Il en résulte une protéine plus légère (48% d'ApoB100 avec 2152 acides aminés) chargée du transport des lipides dans les intestins à la surface des chylomicrons (lipoprotéines se formant lors de la digestion).

A.2.6. Les modifications réversibles de l'ARNm : l'épitranscriptomique

L'édition des ARN n'est pas le seul mécanisme qui modifie les nucléotides de l'ARNm. L'épitranscriptomique désigne toutes les modifications réversibles de l'ARN qui n'altèrent pas les bases (A, C,

G et U). 109 modifications des nucléotides d'ARN ont été recensées (Machnicka *et al.*, 2013) dont 13 pour les ARNm eucaryotes (Cantara *et al.*, 2011). Ces modifications apportent des variabilités supplémentaires au code génétique, notamment en augmentant la masse du nucléotide ciblé (Li et Mason, 2014).

La méthylation [2] de l'ARN est la modification épitranscriptomique de l'ARNm la plus connue et étudiée (Shen *et al.*, 2014; Meyer *et al.*, 2012). Toutes les bases peuvent être méthylées, l'adénosine et la guanosine peuvent porter jusqu'à 3 méthylations et l'uracile jusqu'à deux méthylations. La m^6A (N^6-méthyladénosine) est la modification la plus abondante des ARNm des mammifères (Dominissini, 2012). Les m^6A se retrouvent dans toutes les zones du transcrit mais ils ont une présence plus enrichie sur la fin de la partie codante des gènes et en 5'UTR pour certains tissus. Cette méthylation affecte la stabilisation de l'ARNm (Wang, 2014), la traduction et la localisation de la protéine (Saletore *et al.*, 2013) ou encore l'épissage alternatif (Dominissini, 2012). Elle facilite aussi le transport de l'ARNm dans le noyau (Zheng *et al.*, 2013) dont le mécanisme principal est détaillé ci-après.

A.3. Le transport de l'ARNm : du noyau au cytoplasme

L'ARNm est ensuite transporté dans le cytoplasme via la membrane du noyau. C'est une étape clé de la régulation car la traduction de l'ARNm en protéine se produit dans le cytoplasme. Seulement 5% des ARN messagers quitteront le noyau, le reste sera dégradé. Cheng *et al.* (2006) ont étudié les mécanismes d'exportation de l'ARNm en dehors du noyau. Ils ont montré que l'exportation était dépendante de la coiffe et de l'épissage. Tout d'abord, la CBP80 (Cap-Binding Protein 80) est lié à la coiffe puis un EJC (Exon Junction Complex) est lié à chaque jonction exon-exon durant l'épissage permettant alors au complexe TREX (Transcription Export) de se lier près de la coiffe en 5' (Figure 5). Les ARNm sont alors exportés dans le sens 5' vers 3' grâce au complexe TREX qui va ensuite interagir avec Tapp15, un récepteur des pores nucléaires chargé de l'export des ARNm. Les ARNm vont ainsi être exportés du noyau vers le cytoplasme.

Les ARNm sont alors pris en charge par des protéines de transport qui se fixent à leur queue poly-A et les conduisent le long des microtubules vers leur compartiment cytoplasmique cible (Di Liegro *et al.*, 2014). Ils sont alors maintenus par des microfilaments d'actine pour y être traduits ou sont stockés en réserve en étant liés à des RBP (RNA Binding Protein) formant ainsi le complexe mRNP (messenger RiboNucleoProtein).

2. Ajout d'un groupement méthyle (-CH_3)

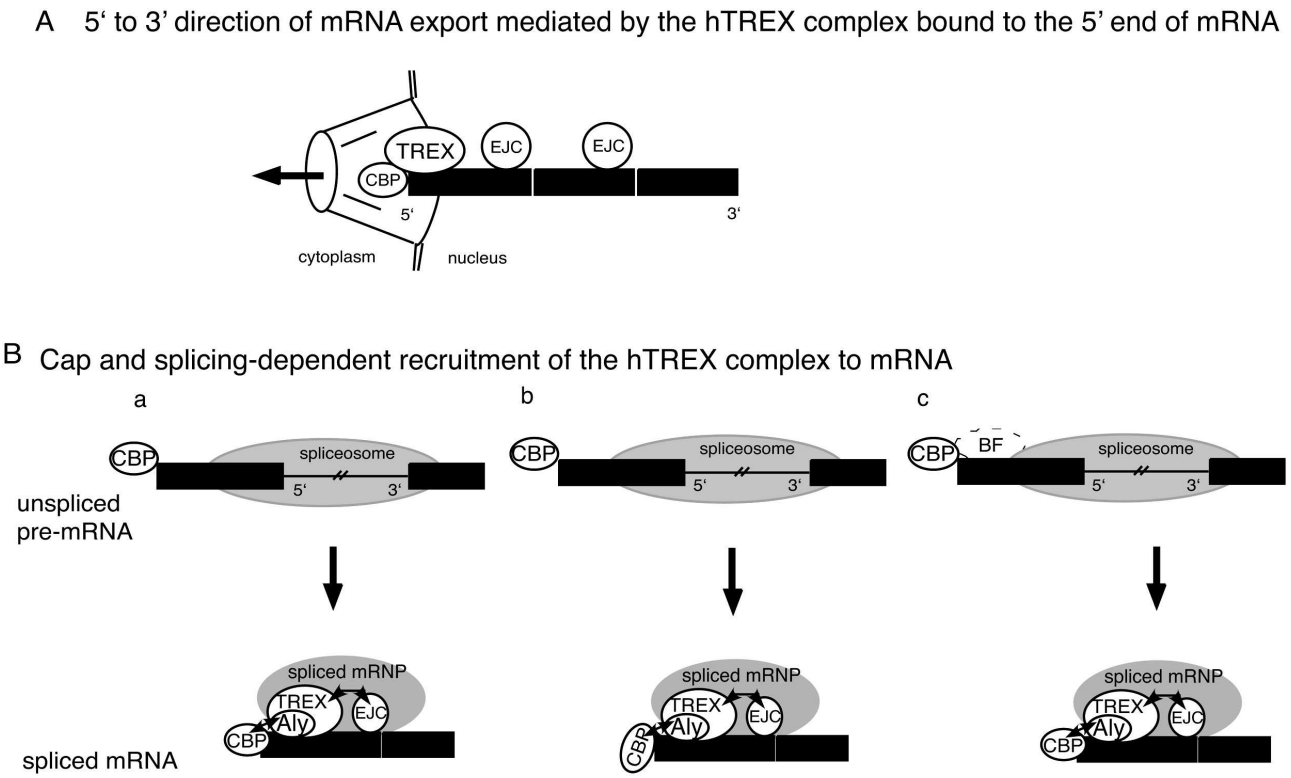

FIGURE 5 – Modèle de l'exportation de l'ARNm par Cheng *et al.* (2006).

A.4. La traduction : la synthèse de la protéine

C'est la biosynthèse d'une protéine à partir de l'ARNm. Comme pour la transcription, on retrouve une phase d'initiation, une phase d'élongation et une phase de terminaison auxquelles on ajoute une phase de recyclage. La traduction nécessite 3 types d'ARN : les ARNr (ARN ribosomal), les ARNt (ARN de transfert) et l'ARNm.

A.4.1. L'initiation

L'initiation consiste en l'assemblage du ribosome 80S avec l'anticodon de l'ARNt initiateur (Met-tRNA) sur le site P du complexe d'initiation 48S (Kapp et Lorsch, 2004). C'est la phase la plus régulée de la synthèse de la protéine et elle nécessite l'intervention d'au moins neuf facteurs d'initiation eucaryotes (les eIFs – eukaryotic initiation factors). Dans la Figure 6, on peut voir les différentes étapes de l'initiation chez les eucaryotes (l'étape 1, le recyclage du ribosome, est détaillée en page 15). L'initiation se déroule en 8 étapes (étapes 2 à 9). Tout d'abord (étape 2), eIF2 lié à de la GTP se lie au Met-ARNt$^{\text{met}}_{i}$ (l'unique ARNt initiateur amino-acétylé avec la méthionine) formant le complexe ternaire eIF2.

Ce complexe se fixe à la sous-unité ribosomique 40S (composée de l'ARNr 18S, d'eIF3, d'eIFA lié au site A et d'eIF1 lié au site E) à l'aide d'eIF5 formant ainsi le complexe de pré-initiation 43S (étape

3). En parallèle, le complexe eIF4F va se fixer sur la coiffe de l'ARNm déroulant la région proximale de la coiffe, aidé par une réaction ATP-dépendante avec eIF4B et la queue poly-A de l'ARNm va se fixer sur le site PABP (Poly-A Binding Protein) d'eIF4GI. Ces fixations vont permettre l'activation de l'ARNm (étape 4) en le stabilisant. Une fois activé, le complexe 43S va s'attacher à la région proximale de la coiffe de l'ARNm en 5' (étape 5). Le complexe 43S va alors se déplacer le long de l'ARNm de 5' vers 3' afin d'identifier le codon d'initiation (étape 6), durant ce déplacement, eIF4E va se détacher du complexe eIF4F en restant fixé à la coiffe. Pour assurer la fiabilité de l'initiation, le complexe 43S doit posséder un mécanisme discriminant les fixations partielles des bases du triplet avec le Met-ARNt$^{\text{met}}_{i}$. L'AUG détecté doit être dans la région consensus suivante : GCC(AG)CCAUGG, on parle d'une initiation dans un contexte fort. C'est eIF1 sur le site E qui va assurer cette fiabilité en discriminant les autres configurations. Lorsque le site d'initiation est reconnu, la fixation va être assurée par l'hydrolyse du complexe binaire eIF2-GTP (étape 7) formant le complexe d'initiation 48S. La sous-unité ribosomique 60S va être liée par l'intermédiaire d'EIF5B couplé à de la GTP qui va entraîner la libération des autres facteurs d'initiation à l'exception d'eIF1A (étape 8). Le départ d'eIF4G va libérer la queue poly-A ce qui facilitera la phase d'élongation. Pour finir l'initiation, la GTP lié à eIF5B est hydrolysée permettant la libération d'eIF1A et d'eIF5B (étape 9). Le complexe d'initiation 80S ainsi formé est alors prêt pour la phase d'élongation.

A.4.2. Initiation alternative

Dans de rares cas, le ribosome est recruté directement au niveau du codon d'initiation, on parle d'une initiation indépendante de la coiffe. Ce mécanisme est permis grâce aux IRES (Internal Ribosome Entry Site) : des régions structurées de l'ARNm (Vagner *et al.*, 2001). Ce mécanisme est utilisé par les cellules eucaryotes notamment dans les situations de stress cellulaire (Wilker *et al.*, 2007). Il est parfois détourné par certains virus (Lytle *et al.*, 2002; Locker *et al.*, 2011).

A.4.3. L'élongation

Durant l'élongation, les acides aminés correspondant aux codons de l'ARNm sont ajoutés progressivement par l'intermédiaire des ARNt. L'élongation se déroule en 3 étapes : incorporation de l'acide aminé apparié, formation du lien peptidique entre l'acide aminé et le reste de la chaîne et translocation du ribosome de 3 nucléotides (Figure 7).

L'aminoacyl-ARNt (l'ARNt avec son acide-aminé correspondant) est apporté au site A vacant par l'intermédiaire d'un facteur d'élongation (eEF1A : eukaryotic Elongation Factor 1A) couplé à une GTP. Ce complexe ternaire peut se fixer au site A, même s'il ne s'agit pas de l'acide aminé apparié.

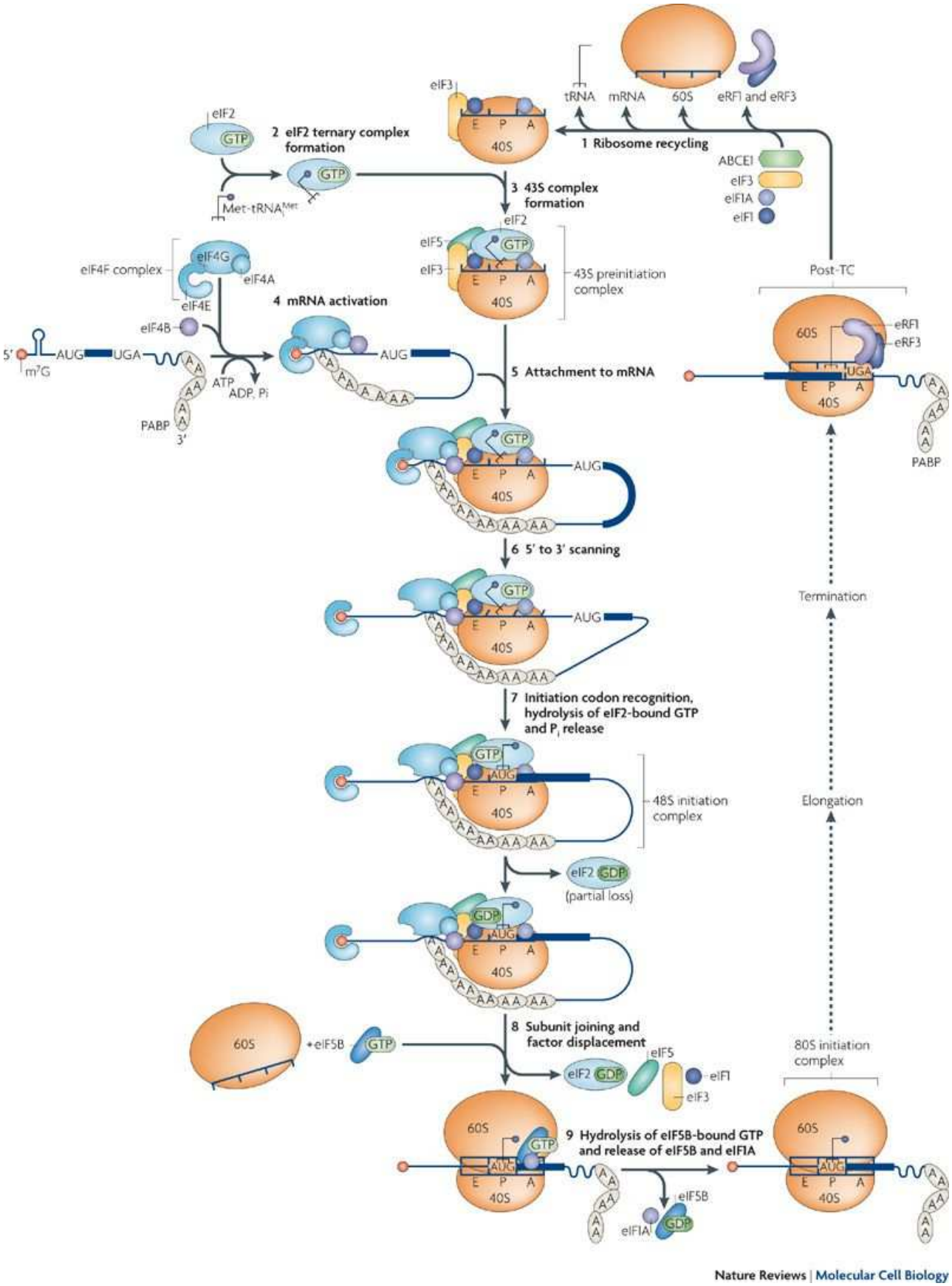

FIGURE 6 – Modèle canonique de la voie de l'initiation de la traduction chez les eucaryotes par Jackson *et al.* (2010)

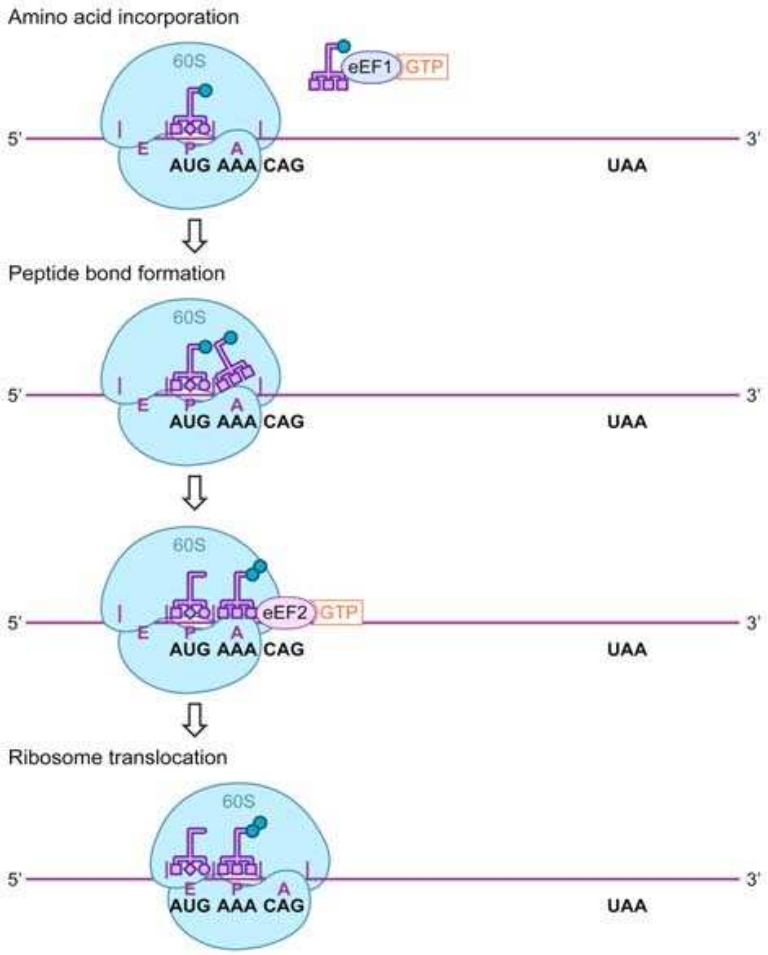

FIGURE 7 – Phase d'élongation de la traduction chez les Eucaryotes par Fritz et Boris-Lawrie (2015).

Toutefois, la suite de l'élongation n'aura lieu que si le bon appariement codon-anticodon a lieu (Table 2). A ce moment-là, la GTP liée à eEF1 est hydrolysée, permettant la fixation de l'aminoacyl-ARNt et le détachement du complexe eEF1-GDP. Une enzyme de la sous-unité ribosomique 60S (la peptidyltransférase) catalyse la liaison entre l'acide-aminé arrivant et la méthionine du codon d'initiation pour la première élongation puis pour les suivantes au dernier acide aminé de la chaîne peptidique se formant au fur et à mesure. L'ARNt sur le site A récupère ainsi la chaîne peptidique. Ensuite eEF2 couple à de la GTP active la translocation du ribosome de 3 nucléotides libérant ainsi le site A en décalant le peptidyl-ARNt (ARNt portant la chaîne peptidique) sur le site P. L'ARNt libéré de son acide-aminé se retrouve alors sur le site E permettant ainsi son relargage. Cette phase d'élongation est répétée jusqu'à l'apparition d'un codon stop qui annonce la phase de terminaison.

A.4.4. La terminaison

La terminaison est la phase d'arrêt de la traduction qui permet le détachement de la protéine nouvellement créée. Lorsqu'un codon UAA, UAG ou UGA arrive sur le site A du ribosome, un facteur

1ère base	2ème base				3ème base
	U	C	A	G	
U	Phénylalanine (Phe/F)	Sérine (Ser/S)	Tyrosine (Tyr/Y)	Cystéine (Cys/C)	U
					C
	Leucine (Leu/L)		Stop (ocre)	Stop (Opale)	A
			Stop (ambre)	Tryptophane (Trp/W)	G
C		Proline (Pro/P)	Histidine (His/H)	Arginine (Arg/R)	U
					C
			Glutamine (Gln/Q)		A
					G
A	Isoleucine (Ile/I)	Thréonine (Thr/T)	Asparagine (Asn/N)	Sérine (Ser/S)	U
					C
	Méthionine* (Met/M)		Lysine (Lys/K)	Arginine (Arg/R)	A
					G
G	Valine (Val/V)	Alanine (Ala/A)	Acide aspartique (Asp/D)	Glycine (Gly/G)	U
					C
			Acide glutamique (Glu/E)		A
					G

TABLE 2 – Table des codons ARN. *codon d'initiation

de terminaison de classe 1 (eRF1 – eukaryotic Release Factor 1) couplé à un eRF3 va reconnaître ce codon comme un codon stop (Table 2). Il va ensuite encourager l'hydrolyse du peptidyl-ARNt par une peptidyltransférase, ce qui va libérer la protéine. Le ribosome va ensuite passer en phase de recyclage.

A.4.5. Le recyclage

Le recyclage est la phase de libération du ribosome afin de le rendre à nouveau disponible pour une prochaine traduction (Figure 6, étape 1). Le ribosome est encore fixé à l'ARNm avec l'ARNt sur le site P qui portait la protéine. La séparation du ribosome 80S pour libérer la sous-unité 60S est catalysée par ABCE1 (ATP-binding cassette sub-family E member 1). La sous-unité 40S est ensuite dissociée de l'ARNm et l'ARNt par l'intermédiaire d'eIF (eIF2D notamment). L'interruption du recyclage peut enclencher une nouvelle traduction après le codon stop.

A.4.6. Polysome

L'ARNm peut être traduit simultanément par plusieurs ribosomes ce qui forme un polysome (ou polyribosome). Cette formation permet d'améliorer l'efficacité de la traduction des ARNm qui ont une durée de vie courte contrairement aux protéines (Rahman et Sadygov, 2017).

A.5. Dégradation de l'ARNm

Des ARNm non-sens (ARNm avec un codon stop prématuré) vont être dégradés, c'est une sorte de contrôle qualité des ARNm. Un ARNm non protégé a une durée de vie de quelques minutes à

quelques heures avant d'être dégradé. En effet, s'il est bien protégé par la cellule, l'ARNm peut durer plusieurs mois. Par exemple chez les mammifères, les globules rouges qui ont éjecté leur noyau continuent de produire l'hémoglobine en conservant les ARNm liés à sa production (Kabanova *et al.*, 2009). Toutefois, la queue poly-A en 3' des ARNm se dégrade à chaque fois que l'ARN est transcrit, ils finiront donc par être quand même dégradés, ce qui explique la durée de vie d'une centaine de jours des érythrocytes.

La stabilisation des ARNm peut être assurée< par des protéines (voir chapitre A.3, page 10) et par des hormones. Par exemple, la prolactine va permettre la stabilisation de l'ARNm produisant la caséine afin d'améliorer la production laitière en augmentant le nombre de fois où l'ARN va être traduit (Choi *et al.*, 2004).

A.6. Modification post-traductionnelle : activation de la protéine

La protéine va être modifiée pour l'adapter à des usages précis, ces modifications peuvent avoir lieu tout au long de la vie cellulaire de la protéine. Ces modifications permettent des modifications d'activité de la protéine (inhibition/activation) et sont souvent catalysées par des enzymes. On en répertorie plusieurs dizaines, voici les plus fréquentes (Khoury *et al.* (2011)) par ordre décroissant.

A.6.1. Clivage de la chaine polypeptidique

Une partie de la séquence de la protéine va être clivée par une protéase qui va hydrolyser la liaison peptidique. Les clivages les plus fréquents se trouvent aux extrémités de la protéine (C-terminale et N-terminale) mais ils peuvent aussi se produire en son sein. Cela permet d'activer des enzymes en clivant un proenzyme (ou zymogène) : un précurseur inactif (Berg *et al.*, 2002). Par exemple, le fibrinogène (une protéine du plasma sanguine) est transformé en fibrine par l'action de la thrombine (une protéase à sérine). La fibrine qui est insoluble dans le plasma sanguin va permettre la formation d'un caillot permettant la coagulation (Mosesson, 2005). La thrombine est elle-même activée lorsque la prothrombine est clivée lors de la cascade de la coagulation.

A.6.2. Formation des ponts disulfures

Il s'agit d'une liaison covalente entre les atomes de soufre de deux cystéines d'une protéine. Cette liaison peut s'effectuer entre n'importe quelle cystéine et plusieurs ponts peuvent être créés. Actuellement, il est difficile de prédire comment les ponts vont se former à partir de la séquence protéique. Ces liaisons peuvent jouer plusieurs rôles : stabiliser la structure ou maintenir une liaison

entre différentes chaînes peptidiques ou sous-unités d'une même chaîne. C'est le cas par exemple de l'insuline dont les extrémités sont clivées puis fixées par 3 ponts disulfure avant de cliver la partie joignant les deux extrémités (Yuan *et al.*, 1999).

A.6.3. Phosphorylation

Il s'agit de l'ajout d'un phosphate sur la sérine, la tyrosine, l'histidine ou la thréonine par une kinase. C'est le mécanisme de régulation le plus fréquent des protéines (Khoury :2011). L'ajout d'un groupement phosphate qui est chargé négativement va changer la configuration de la protéine qui peut être ainsi activée ou désactivée. Le mécanisme inverse est la déphosphorylation, catalysée par une phosphatase. Ces deux mécanismes permettent une inactivation/activation des protéines comparable à un interrupteur. Par exemple, la cascade des MAPK (Mitogen-Activated Protein Kinase) est impliquée dans la réponse aux facteurs de croissance nécessaires à l'induction de la mitose (Pearson *et al.*, 2001). Les MAPK sont inactives quand elles sont déphosphorylées, la phosphorylation par des kinases précédentes les active ce qui leur permet d'activer les kinases suivantes.

A.6.4. Acétylation

Il s'agit de l'ajout d'un groupe acétyle sur l'extrémité N-terminal de la protéine (appelée N-acétylation) ou sur une lysine. La N-acétylation est catalysée par des N-terminal acétyltransférases (NATs) qui vont transférer le groupe acétyle d'un acétyl-CoA (acétyl-coenzyme A, une forme activée de l'acide acétique). 85% des protéines humaines sont N-acétylées (Van Damme *et al.*, 2011). L'acétylation joue un rôle sur l'activation des protéines en neutralisant la charge positive des lysines. Par exemple, l'acétylation des histones favorise la transcription. En effet, les queues des histones interagissent de manière maximale avec l'ADN grâce aux charges positives des lysines. Quand les charges des lysines sont neutralisées par les HAT (Histone AcétylTransférase) en les acétylant, l'histone libère un peu l'ADN ce qui augmente l'activité du promoteur (Stasevich *et al.*, 2014).

A.6.5. Glycosylation

Il s'agit de l'ajout d'un glucide à une asparagine (N-glycosylation), à une thréonine ou à une sérine (O-glycosylation). La N-glycosylation (3ème modification la plus fréquente) et l'O-glycosylation (7ème plus fréquente) sont généralement catalysées par des glycosyltransfrérases. Les glycosylations jouent un rôle important dans la protection de la protéine contre la protéolyse et concernent essentiellement les protéines membranaires ce qui assure la stabilité de la cellule. Les glycosylations

sont aussi présentes dans les protéines sécrétées où elles permettent leur transport et leur adressage (Schwarz et Aebi, 2011; Van den Steen *et al.*, 1998).

A.6.6. L'amidation en C-terminal

Elle consiste en un clivage de la protéine en C-terminal qui ajoute un groupement amide créant ainsi une glycine. L'amidation diminue la charge électrique de la protéine et augmente sa polarité ce qui protège la protéine d'une protéolyse. Cette réaction est catalysée par une Peptidylglycine alpha-amidating monooxygenase (PAM - Kolhekar *et al.*, 1997; Eipper *et al.*, 1993).

A.6.7. Hydroxylation

L'hydroxylation est l'ajout d'un groupe hydroxyle (-OH) sur une lysine ou une proline qui est catalysée par une hydroxylase (prolyl hydroxylase ou lysil hydroxylase). Pour être hydroxylé, il faut que l'acide aminé (proline ou lysine) soit suivi d'une glycine et cette réaction nécessite la présence d'acide ascorbique (ou vitamine C). L'hydroxylation joue un rôle dans la stabilisation d'une protéine en la protégeant des protéases et dans sa réticulation (création de liens entre différents polymères formant ainsi un réseau tridimensionnel) en permettant la formation de liaisons covalentes entre différentes chaînes polypeptidiques. Par exemple, l'hydroxylation des prolines et des lysines du tropocollagène permet à cette protéine de s'assembler à d'autres tropocollagènes pour former le collagène (Anttinen *et al.*, 1981). C'est l'absence de Vitamine C empêchant l'hydroxylation des tropocollagènes (donc du collagène) qui est à l'origine du scorbut.

A.6.8. Méthylation

La Méthylation est l'ajout d'un groupement méthyl sur un azote d'une lysine ou d'une arginine. La Lysine peut être méthylée une, deux ou trois fois par des lysine méthyltransférases et l'Arginine peut être méthylée une ou deux fois par des PRMTs (Protein Arginine MethylTransférases). La méthylation est très étudiée chez les histones car elle joue un rôle prépondérant dans la régulation de l'hétérochromatine (la forme condensée de l'ADN enroulé autour des histones). L'histone méthyltransférase (HMT) va méthyler les lysines présentes sur la queue N-terminale de l'histone permettant la formation de la chromatine (Bártová *et al.*, 2008).

A.6.9. Ubiquitination

Il s'agit des étapes amenant à la liaison covalente de la glycine en C-terminale de l'ubiquitine (un peptide de 76 acides aminés qui est très conservé chez les eucaryotes) sur une ou plusieurs lysines (bien que dans des cas rares la fixation peut se faire sur une sérine, une cystéine ou une thréonine). La présence de 4 ubiquitines sur une protéine va permettre la reconnaissance puis la dégradation de cette protéine par le complexe protéolytique du protéasome (Hershko et Ciechanover, 1998). En revanche, la mono- et la di-ubiquitination assure la stabilité de la protéine ciblée, son activation et joue un rôle dans la réparation de l'ADN comme par exemple sur la protéine Ras (Jura *et al.*, 2006). Il existe un autre mécanisme proche de l'ubiquitination : la SUMOylation. Comme pour l'ubiquitination, la protéine SUMO (small ubiquitin modifier of proteins) est lié sur une lysine de la protéine ciblée. Dans certains cas, la SUMOylation est antagoniste au mécanisme de dégradation guidé par l'ubiquitine (Ramachandran *et al.*, 2015).

En plus de ces modifications les plus fréquentes, ils en existent d'autres qui en général n'affectent qu'une seule base précise en y ajoutant un élément chimique par l'intermédiaire d'une enzyme (iodation ou sulfonation de la tyrosine, succinylation de la lysine ou l'ajout d'un lipide : l'ancrage lipidique). Les modifications les plus communes de chaque acide-aminé est présenté dans la Table 3, on constate que la leucine, l'isoleucine et la phénylalanine ne sont pas modifiées dans les connaissances actuelles.

Code	Acide aminé	Modifications
A	Alanine	N-acétylation
C	Cystéine	Ponts disulfures, oxydation, ancrage lipidique, N-acétylation, S-nitrosylation
D	Acide aspartique	Isomérisation en isoaspartate
E	Acide glutamique	Cyclisation en acide pyroglutamique, gamma-carboxylation
G	Glycine	Ancrage lipidique, N-acétylation
H	Histidine	Phosphorylation
K	Lysine	Acétylation, ubiquitination, SUMOylation, méthylation, hydroxylation
M	Méthionine	N-acétylation, oxydation
N	Asparagine	Désamidation en aspartate ou en isoaspartate, N-glycosylation
P	Proline	Hydroxylation
Q	Glutamine	Cyclisation en acide pyroglutamique, désamidation en acide glutamique
R	Arginine	Déimination en citrulline, méthylation
S	Sérine	Phosphorylation, O-glycosylation, N-acétylation
T	Thréonine	Phosphorylation, O-glycosylation, N-acétylation
V	Valine	N-acétylation
W	Tryptophane	Oxydation, formation de Kynurénine
Y	Tyrosine	Sulfonation, phosphorylation

TABLE 3 – Modifications post-traductionnelles fréquentes par acide aminé.

A.7. Les microARN : leurs rôles dans l'expression des gènes

Les microARN (ou miARN) sont de petits ARN d'environ 22 nucléotides (en général de 21 à 24 nucléotides) présents uniquement chez les eucaryotes. Les microARN jouent un rôle déterminant dans la régulation en s'appariant à des séquences complémentaires des ARNm entraînant la dégradation de cet ARNm ou inhibant sa traduction. Ils sont très abondants et ciblent environ 60% des gènes (Friedman *et al.*, 2009).

A.7.1. Biogénèse des microARN : la voie canonique

Les gènes microARN ne sont pas directement transcrits en microARN. Comme les ARNm, ils doivent subir des modifications post-transcriptionnelles. De plus, les transcrits non codants des gènes de miARN sont souvent « polycistroniques », c'est-à-dire que plusieurs miARN différents sont générés à partir d'un seul transcrit général. On estime chez l'homme qu'environ 40% de miARN sont ainsi organisés en clusters (Altuvia *et al.*, 2005).

Dans la Figure 8, les étapes de la voie classique de la biosynthèse des miARN chez les métazoaires sont présentées (Wahid *et al.*, 2010). Tout d'abord, le gène du microARN est transcrit par l'ARN polymérase II en un ARN primaire : pri-miARN (primary miARN). Ce pri-miARN est replié sur lui-même formant une tige-boucle imparfaite (une partie de la tige n'est pas 100% complémentaire). Les deux extrémités en dehors de la structure tige-boucle sont alors clivées par un complexe appelé « microprocesseur » pour former le miARN précurseur (pré-miARN). Ce complexe est composé de la protéine Drosha et d'une protéine de liaison aux ARN double-brin : DGCR-8 (DiGeorge syndrome Critical region Gene-8) chez les mammifères. Le pré-miARN d'une taille d'environ 65 nucléotides possède une souche excédentaire en 3' de 2-3 nucléotides qui est reconnu par l'exportine 5 (EXP5). L'EXP5 va alors guider le pré-miARN jusqu'au cytoplasme en permettant sa sortie via les pores nucléaires. Dans le cytoplasme, la ribonucléase III Dicer va interagir avec une protéine possédant un domaine de fixation des ARN double brin : la TRBP (TAR RNA-binding Protein). Cette interaction va permettre à Dicer de reconnaître le pré-miARN pour hydrolyser la boucle du pré-miARN générant ainsi le duplex miARN/miARN. Le duplex va alors être fixé à une protéine de la famille des Argonautes (Ago1-4) et une hélicase. Un des deux miARN de cette structure double-brin va alors être sélectionné, généralement le brin avec l'extrémité 5' la moins stable au niveau de l'appariement (Krol *et al.*, 2010). L'hélicase va alors séparer les deux brins créant ainsi le complexe RISC (RNA-induced silencing complex) composé d'AGO1-4 et du miARN mature d'une taille d'environ 21 nucléotides. Le complexe RISC va alors pouvoir jouer un rôle dans la régulation.

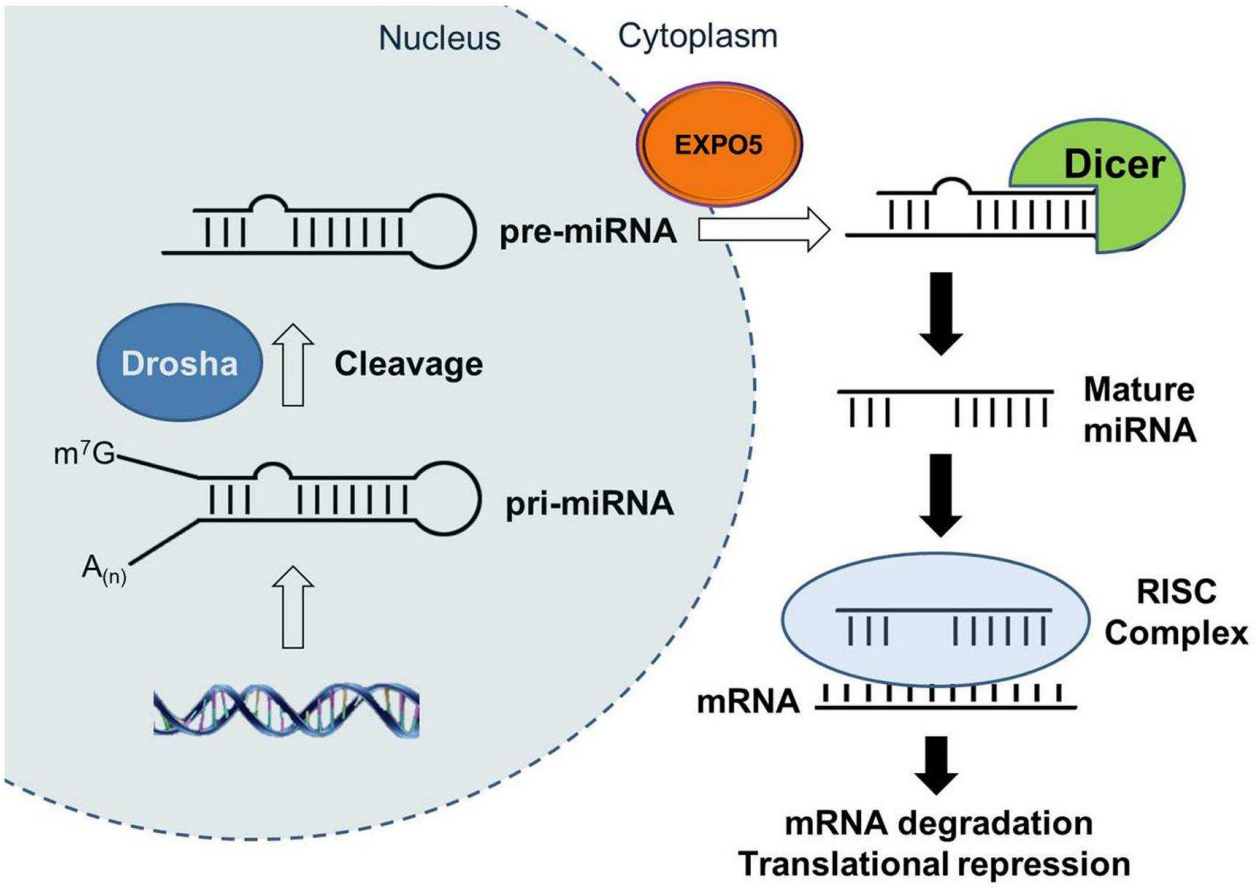

FIGURE 8 – Voie canonique de la biogénèse des microARN par Jung et Suh (2015)

A.7.2. Mode d'action des microARN dans la régulation

Les cibles des miARN se situent dans la région 3'UTR (Untranslated Transcribed Region) des gènes chez les animaux, la complémentarité cible-miARN est presque toujours partielle (dans le cas d'une complémentarité totale, l'ARNm sera coupé par Ago2). Les ARNm contiennent souvent plusieurs cibles pour le même miARN ou pour différents miARN (Pillai *et al.*, 2007). De plus, l'efficacité inhibitrice du miARN est corrélée avec le nombre de cibles présentes dans la région 3'UTR de l'ARNm (Broderick *et al.*, 2011).

Les microRNAs inhibent l'expression des gènes ciblés soit en encourageant la dégradation des ARNm ciblés soit en réprimant la traduction de ces ARNm. Lorsque l'appariement est parfait et que la protéine Ago2 est dans le complexe RISC, cette dernière va lyser l'ARNm entraînant sa dégradation (Meister *et al.*, 2004). Une déadénylation des ARNm ciblés par des miARN peut aussi entraîner à la dégradation de l'ARNm. Les ARNm déadénylés sont ensuite décoiffés et dégradés (Behm-Ansmant *et al.*, 2006).

Ils existent un grand nombre de mécanismes pouvant induire l'inhibition de la traduction par des microARN (Pillai *et al.*, 2007; Gu et Kay, 2010). La traduction peut être inhibée au niveau

de l'initiation soit par compétition avec le facteur d'initiation eIF4G, soit en empêchant la jonction du ribosome 60S avec le complexe d'initiation 48S. D'autres voies impliquent un ralentissement des ribosomes, diminuant l'efficacité de la phase d'élongation de la protéine. La traduction peut aussi être inhibée avec une terminaison prématurée de la traduction entraînant une dégradation de la protéine. Les microARN vont ainsi ralentir et retarder la traduction en agissant sur ces trois phases, cela pourrait permettre une régulation plus précise en comparaison à la dégradation de l'ARNm (Gu et Kay, 2010).

A.7.3. Biogénèse des microARN : les voies alternatives

La voie canonique présentée au préalable nécessitant les protéines Drosha et DGCR-8 pour le clivage des extrémités et Dicer pour l'hydrolyse de la boucle possède des alternatives. Elles sont classées en 3 catégories (Figure 9) : les Drosha et DCR8-indépendantes (ou microprocesseur-indépendantes – Babiarz *et al.*, 2008), les Dicer-indépendantes et les TUTase-dépendantes. Bien qu'importantes, ces voies alternatives ne représentent qu'un pour cent de la biosynthèse des microARN.

Les mirtrons (Ruby *et al.*, 2007) sont des petits ARN précurseurs présents dans la région intronique d'un ARNm qui sont clivés lors de l'épissage par le splicéosome. Une fois le lasso (formé par l'épissage) détaché, la séquence est modelé en une structure tige-boucle semblable à celle des pré-miARN contournant ainsi le clivage par Drosha. Certains mirtrons possèdent des extrémités 5' ou 3' qui nécessitent d'être tronquées avant d'être clivés par Dicer (Wen *et al.*, 2015). Dans les voies microprocesseur-indépendantes, on retrouve aussi les miARN avec une coiffe 7-méthylguanosine(m7G) qui sont générés directement via la transcription et exportés par l'exportine 1 ; d'autres ARN non codants comme les snoARN (Taft *et al.*, 2009) et les ARNt peuvent aussi être clivés pour produire un pré-miARN.

Les miARN primaires de groupe II TUTase (Terminal Uridylyl Transférases) dépendant sont produits en miARN précurseur avec une partie 3' excédentaire plus courte (d'un nucléotide seulement). Une Uridine doit être ajoutée par les TUTase (TUT2, TUT4 et TUT7) pour que l'activité de Dicer soit efficace (Aphasizhev *et al.*, 2002).

La voie des miARN indépendante de Dicer consiste en un attachement direct du pré-miARN sur Ago-2 sans clivage préalable par Dicer. Ce cas rare a été observé avec miR-451 qui s'incorpore directement dans les protéines Ago-2 en raison de sa longueur courte. La protéine Ago-2 va alors cliver la partie en position de brin passager et des co-facteurs vont raccourcir le miARN par l'extrémité 3' (Treiber *et al.*, 2018).

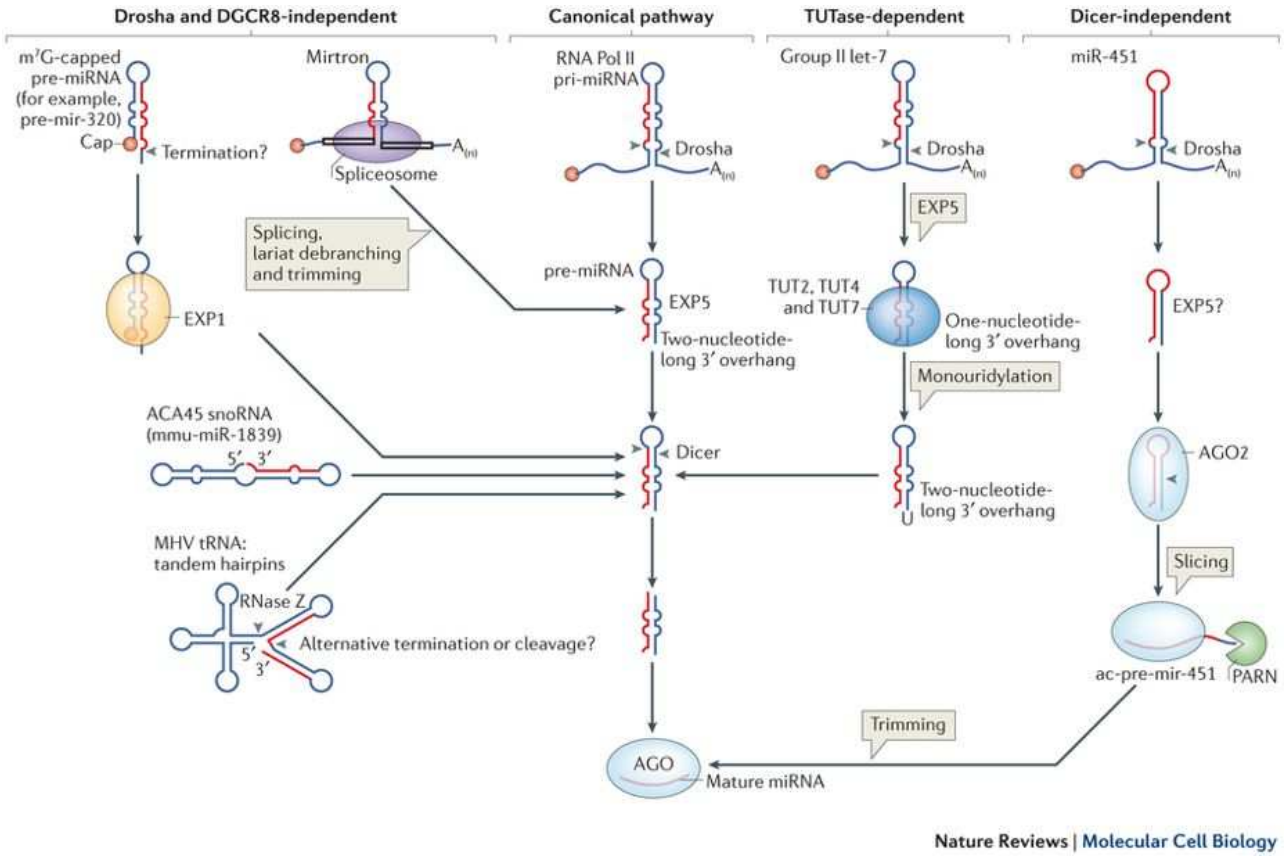

FIGURE 9 – Voies alternatives de la biogènese des microRNAS par Ha et Narry Kim (2014)

A.8. Les longs ARN non codants : des ARN très polyvalents

Comme on l'a constaté dans les chapitres précédents, tous les ARN ne sont pas codants (ARNt, ARNr, miARN notamment) mais ils jouent principalement un rôle dans la traduction ou l'inactivation/dégradation de l'ARNm. Le rôle des ARN ne codant pas pour des protéines a tendance à être sous-estimé (Amaral *et al.*, 2008) alors qu'une étude récente a montré que 20% des gènes annotés comme codants chez l'homme seraient en réalité non codants (Abascal *et al.*, 2018). En plus, des ARN non codants, il existe les long ARN non codants (ou lncRNA : long non coding RNA). Les lncRNA sont définis arbitrairement comme étant des ARN non codants d'une taille supérieure à 200 nucléotides. Ils peuvent être retrouvés au sein des gènes soit dans les introns, soit antisens, soit chevauchant les exons ou alors être situés entre des gènes. Dans ce cas-là on parle de lincRNA (long intergenic non coding RNA) (Ma *et al.*, 2013).

Les lncRNA sont généralement faiblement exprimés (Derrien *et al.*, 2012; Cabili *et al.*, 2011) mais malgré cela ils jouent un rôle significatif dans de nombreux processus (Ponting *et al.*, 2009) : notamment l'épissage et la transcription (Geisler et Coller, 2013), la traduction (Carrieri *et al.*, 2012), l'empreinte (Brockdorff *et al.*, 1991; Gupta *et al.*, 2010), la pluripotence des cellules souches (Yu *et al.*, 2018), le cycle cellulaire (Yang *et al.*, 2012) et l'apoptose (Han *et al.*, 2014). Ils sont

aussi impliqués dans le développement de cancers (Bhan *et al.*, 2017).

Un des lncRNA le plus connu est codé par le gène *Xist* (X inhibitory specific transcript) et il est impliqué dans l'inactivation du X chez les femelles mammifères. L'inactivation du X est un processus entraînant l'inactivation d'un des chromosomes X au hasard (ou le chromosome X paternel dans le cas des marsupiaux – Cooper *et al.*, 1971) durant les premières étapes du développement embryonnaire. Le lncRNA *Xist* est transcrit en grande quantité et va se fixer sur le chromosome X dont il est issu pour l'inactiver (Brockdorff *et al.*, 1991). La transcription de *Xist* par un des deux chromosomes X étant inactivée par la méthylation de son promoteur (Carrel et Willard, 2005). Cette inactivation de *Xist* est ensuite stabilisée de manière plus pérenne par des modifications épigénétiques.

A.9. Rôle de l'épigénétique dans l'expression des gènes

L'épigénétique désigne les mécanismes modifiant l'expression des gènes sans altérer leur séquence d'ADN ; ces modifications sont transmissibles (lors des divisions cellulaires et donc potentiellement transmises à la descendance) et réversibles (Dupont *et al.*, 2009). Les modifications épigénétiques sont induites par l'environnement de la cellule et permettent d'expliquer la différenciation cellulaire. En effet, presque toutes nos cellules contiennent la même séquence d'ADN et c'est lors de la morphogénèse (développement de l'œuf fécondé en embryon) que les cellules souches totipotentes vont se spécialiser en des lignées cellulaires pluripotentes pour ensuite acquérir un type cellulaire définitif (neurone, cellule musculaire, *etc.*). Cette différenciation se traduit par une différence d'expression des gènes certains étant activés et d'autres inhibés et elle est conservée grâce à des mécanismes épigénétiques (Reik, 2007).

Les modifications épigénétiques sont principalement des marques biochimiques (méthylation, acétylation, *etc.*) ajoutées par des enzymes spécifiques soit directement sur l'ADN, soit sur les histones.

A.9.1. La structure de la chromatine et état des histones

Pour comprendre l'intérêt des histones, il est important de souligner que les 46 chromosomes d'une cellule humaine représentent 2 mètres d'ADN et qu'ils doivent être contenus dans la cellule qui mesure entre 10 et 100 μm de diamètre. Les histones jouent alors un rôle important dans la condensation de l'ADN. Les différentes étapes de cette condensation sont détaillées dans la Figure 10. Le nucléosome se compose de deux histones de chaque classe (H2A, H2B, H3 et H4) associé en un octamère. L'ADN double brin s'enroule d'abord autour de ces nucléosomes sur environ 150 paires de bases puis une histone de classe H1 stabilise le nucléosome en liant les deux brins d'ADN sortant

du cœur du nucléosome. Les nucléosomes s'enroulent ensuite sur eux-mêmes formant ainsi des fibres de chromatine qui vont s'enrouler de plus en plus jusqu'à former la chromatide. Les chromosomes sont composés de deux chromatides complètement identiques reliées par le centromère.

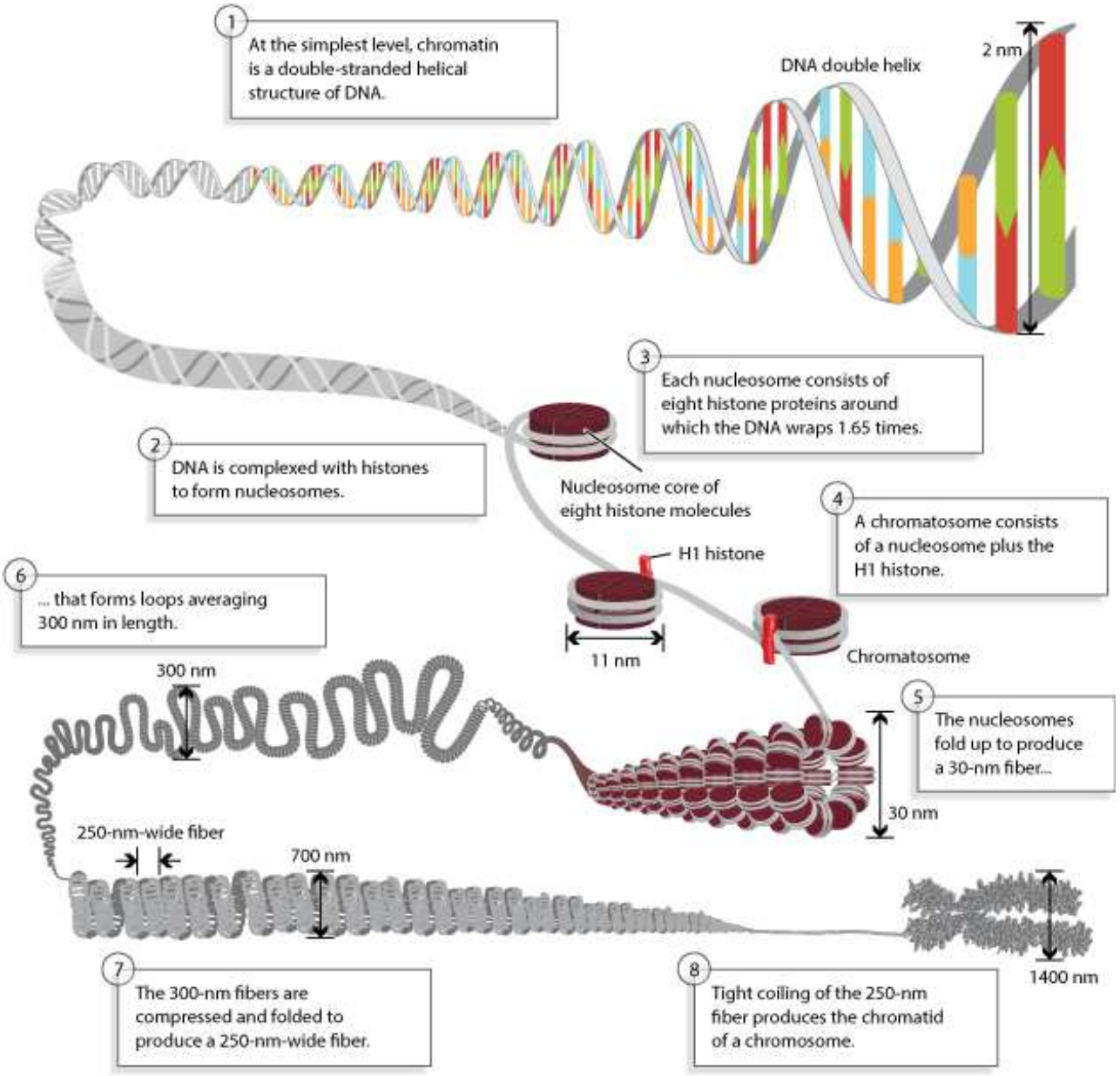

FIGURE 10 – Les différents niveaux de condensation de l'ADN par Annunziato (2008)

Les fibres de chromatine peuvent être plus ou moins denses. Quand elles sont très denses, il s'agit d'une hétérochromatine : l'ADN est tellement condensé que les ARN polymérases ne peuvent pas accéder aux gènes et ces derniers ne sont donc pas exprimés. A l'inverse, quand elles sont moins condensées (euchromatine), les gènes peuvent êtres exprimés. Le passage d'une densité à l'autre est contrôlé par les modifications épigénétiques des histones. Il existe toutefois des hétérochromatines constitutives (Hsu et Arrighi, 1971), c'est-à-dire qu'elles sont constamment condensées, les princi-

pales sont les centromères et les télomères (régions très répétitives présentes aux extrémités d'un chromosome).

L'ouverture ou la fermeture de l'ADN par modification de la condensation va donc être fortement influencée par l'état des histones. Cet état dépend des modifications post-traductionnelles des histones. On parle de code histone pour désigner ces modifications qui régulent la transcription (Jenuwein et Allis, 2001). L'ouverture est effectuée par l'acétylation de lysine de l'histone (détaillé page 17) par l'intermédiaire des HAT (Stasevich *et al.*, 2014). L'interaction entre les nucléosomes va diminuer passant la chromatine à l'état chromatinien ce qui va permettre l'action des facteurs de transcription et démarrer la transcription. Cette ouverture peut être refermée en retirant l'acétylation par le biais des histones désacétylases (HDAC). L'acétylation joue aussi un rôle important dans la réparation de l'ADN (Gong et Miller, 2013).

Il existe un mécanisme opposé à l'acétylation qui modifie les histones pour conserver la condensation sous forme d'hétérochromatine : il s'agit de la méthylation. De manière générale, la méthylation est antagoniste à l'acétylation car les lysines acétylées doivent d'abord être désacétylées avant d'être méthylées (Rice et Allis, 2001). En effet, la méthylation va agir sur les histones en méthylant les lysines présentes sur les queues des histones (détaillé page 18), ces méthylations favorisent l'enroulement de l'ADN sur les histones (Greer et Shi, 2012). Toutefois, d'autres méthylations peuvent favoriser l'activation de la transcription en fonction du nombre de méthylation des lysines (jusqu'à 3 méthylations par lysine). La mono-méthylation d'une lysine est souvent lié à une activation de la transcription, par exemple sur H3K9[3], H3K27, H3K79 ou H4K20 (Barski *et al.*, 2007).

D'autres modifications post-tradutionnelles des histones peuvent réguler la transcription. La phosphorylation et l'ubiquitination favorisent généralement son activation et la SUMOylation en revanche la réprime. La majorité des modifications des histones découvertes avec le site modifié et les conséquences éventuelles a été recensé par (Cell Signaling Technology, 2018).

A.9.2. La méthylation de l'ADN

Il existe un autre phénomène fréquent d'épigénétique : la méthylation de l'ADN. Elle consiste en l'ajout d'un groupement méthyle sur une cytosine formant une 5-méthylcytosine. Chez les mammifères, la méthylation se fait principalement au sein des îlots CpG (Cytosine-phosphate-Guanine) des régions de l'ADN riches en Guanine et Cytosine (Bird, 1986). Cette réaction est catalysée par des ADN méthyl-transférases (DNMT). La DNMT1 maintient les méthylations présentes lors de chaque division cellulaire en parallèle de la réplication de l'ADN (Bestor, 2000; Robert *et al.*, 2003). La

3. H3K9 signifie que la modification a lieu sur la lysine (K) en position 9 de l'histone 3 (H3).

DNMT2 plus courte se fixe à l'ADN et joue un rôle dans la conservation de la structure et de la fonction des centromères (Dong *et al.*, 2001). La DNMT2 joue aussi un rôle dans la méthylation de l'ARN vu précédemment. DNMT3a et DNMT3b sont impliquées dans les mutations *de novo* (Okano *et al.*, 1999).

Les DNMTs régulent aussi l'expression des gènes en agissant directement ou indirectement sur la transcription. Lorsque la région promotrice d'un gène est méthylée, les facteurs de transcription ne peuvent plus s'y fixer ce qui réprime l'expression de ce gène (Comb et Goodman, 1990; Bell et Felsenfeld, 2000). La méthylation de l'ADN peut aussi favoriser l'action des HDAC permettant la condensation des chromatines (El-Osta et Wolffe, 2001; Irvine *et al.*, 2002).

La méthylation de l'ADN peut aussi jouer un rôle positif sur la transcription en augmentant son activité. Alors que les méthylations du promoteur bloquent l'initiation, celles présentes au sein des gènes peuvent stimuler la phase d'élongation (Jones, 2012). Les méthylations de l'ADN intragénique peuvent aussi moduler l'épissage alternatif (Maunakea *et al.*, 2013) et peuvent être modulées durant la différenciation cellulaire et les cancers (Kulis *et al.*, 2013).

A.9.3. L'empreinte parentale

Les méthylations de l'ADN et des histones sont généralement conservées au cours de la réplication de l'ADN grâce à l'action de la DNMT1 notamment. Toutefois, toutes ces méthylations ne sont pas forcément transmises à la descendance. Avant les premières divisions de la cellule-œuf, les génomes parentaux subissent une forte démythélation mais certains gènes et histones du père ou de la mère vont rester méthylés, c'est l'empreinte parentale (Li *et al.*, 1993; Barlow et Bartolomei, 2014). Les méthylations ainsi conservées sont généralement celles ayant eu lieu pendant la formation des cellules germinales (ovules et spermatozoïdes). Pour le moment, ce phénomène d'empreinte parentale partielle dans le génome semble unique aux thériens (mammifères placentaires et marsupiaux - Reik et Walter, 2001) et aux angiospermes (plantes à fleurs Autran *et al.*, 2005). L'inactivation du chromosome X (voir chapitre A.8, page 23) peut être considérée comme une empreinte parentale touchant l'un des deux chromosomes X dans son intégralité. Les gènes soumis à empreinte avec une des copies inactivées auront alors une expression monoallélique, ce qui peut entraîner dans certains cas des dysfonctionnements (Wilkins et Úbeda, 2011).

B. Étudier les mécanismes de régulation de l'expression des gènes

On a vu que l'expression des gènes était régulée par de nombreux mécanismes, il est crucial de pouvoir les analyser et les identifier. En effet, une cartographie précise de ces mécanismes permet de mieux identifier des causes de maladies, d'optimiser des caractères via la sélection ou encore d'améliorer les connaissances actuelles sur ces mécanismes. Dans les pages suivantes, je vais détailler quelques approches bio-informatiques et expérimentales permettant d'étudier ces mécanismes.

B.1. Séquençages des génomes et des transcriptomes

L'étape primordiale pour une étude bio-informatique de la régulation de l'expression des gènes chez un individu est de séquencer son génome et son transcriptome (l'ensemble des transcrits produits). Avec l'avènement des technologies de séquençage de nouvelle génération (NGS pour Next Generation Sequencing en anglais) en 2007 (Figure 11), un grand nombre de génomes de différentes espèces ont été séquencés et le coût a fortement diminué au-delà des prédictions de la loi de Moore [4].

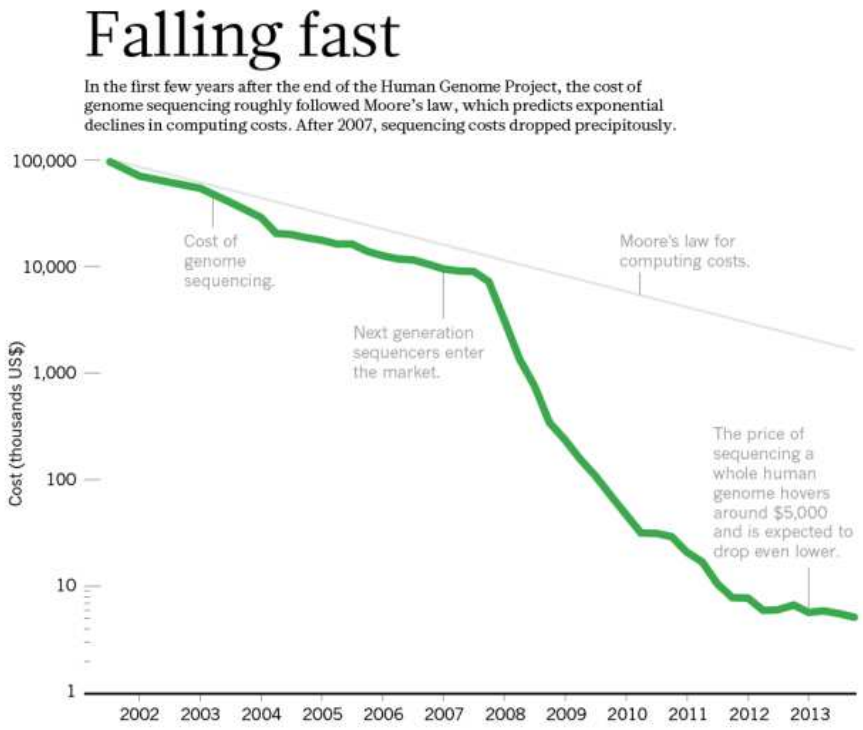

FIGURE 11 – Evolution du coût de séquençage en comparaison à loi de Moore par Hayden (2014).

4. Le nombre de transistors des microprocesseurs double tous les deux ans. Elle a connu de nombreuses adaptations notamment sur la diminution progressive du coût de calcul informatique.

B.1.1. Technologies de séquençage : décrypter l'ADN et l'ARN

Le séquençage Sanger

Le séquençage Sanger est la première technique de séquençage mis au point (Sanger et Coulson, 1975), c'est une méthode classique qui bien que bas débit est encore utilisée de nos jours. Depuis 1975, la méthode a évolué au fil des découvertes (ex : utilisation de marqueurs fluorescents à la place de marqueurs radioactifs). Le principe de fonctionnement de cette méthode repose sur l'initiation de la polymérisation de l'ADN à séquencer à partir d'une amorce à l'aide de l'ADN polymérase I. En plus des 4 désoxyribonucléotides classiques (dATP, dCTP, dGTP et dTTP), on incorpore en faible proportion un des 4 didéoxynucléotides qui sont terminateurs de la chaîne (ddATP, ddCTP, ddGTP et ddTTP). Un traceur fluorescent a été attaché soit aux oligonucléotides, soit aux didésoxyribonucléotides. On obtient alors différents fragments de taille différentes en fonction de l'endroit où le didésoxyribonucléotide a été intégré. Puis, on fait migrer ces différents fragments par électrophorèse. En fonction du didéoxynucléotide intégré, on répère toutes les positions du nucléotide concerné et on reconstitue ainsi la séquence d'ADN.

NGS : le séquençage haut-débit

Pour la génomique et la bio-informatique, l'arrivée du séquençage haut-débit a été une grande révolution biotechnologique car plus rapide et moins chère. En effet, avec la méthode Sanger, il a fallu 13 ans (de 1990 à 2003) pour séquencer le génome humain dans plusieurs laboratoires internationaux et ce projet (Human Genome Project - Chial, 2008) a coûté 3 milliards de dollars. Actuellement, grâce aux nouveaux séquenceurs NGS on peut séquencer le génome humain en quelques jours pour un coût de mille dollars. Il existe 4 méthodes principales de séquençage de l'ADN utilisées en NGS : le pyroséquençage (Roche 454 - Margulies *et al.*, 2005), le séquençage par ligation (SOLiD pour Sequencing by Oligo Ligation Detection - McKernan *et al.*, 2009), le séquençage par synthèse (Illumina, anciennement Solexa - Bentley *et al.*, 2008) et le séquençage par détection des ions H^+ (Ion Torrent : Proton/PGM - Pennisi, 2010).

Ces technologies se déroulent en 3 étapes (cf Figure 12) :

1. La préparation des banques ou librairies de séquençage : la séquence d'ADN est fragmentée de manière aléatoire par la méthode WGS (Whole Genome Shotgun) soit en utilisant la sonication, soit par digestion enzymatique. Des adaptateurs spécifiques sont alors ajoutés par des ADN ligases.

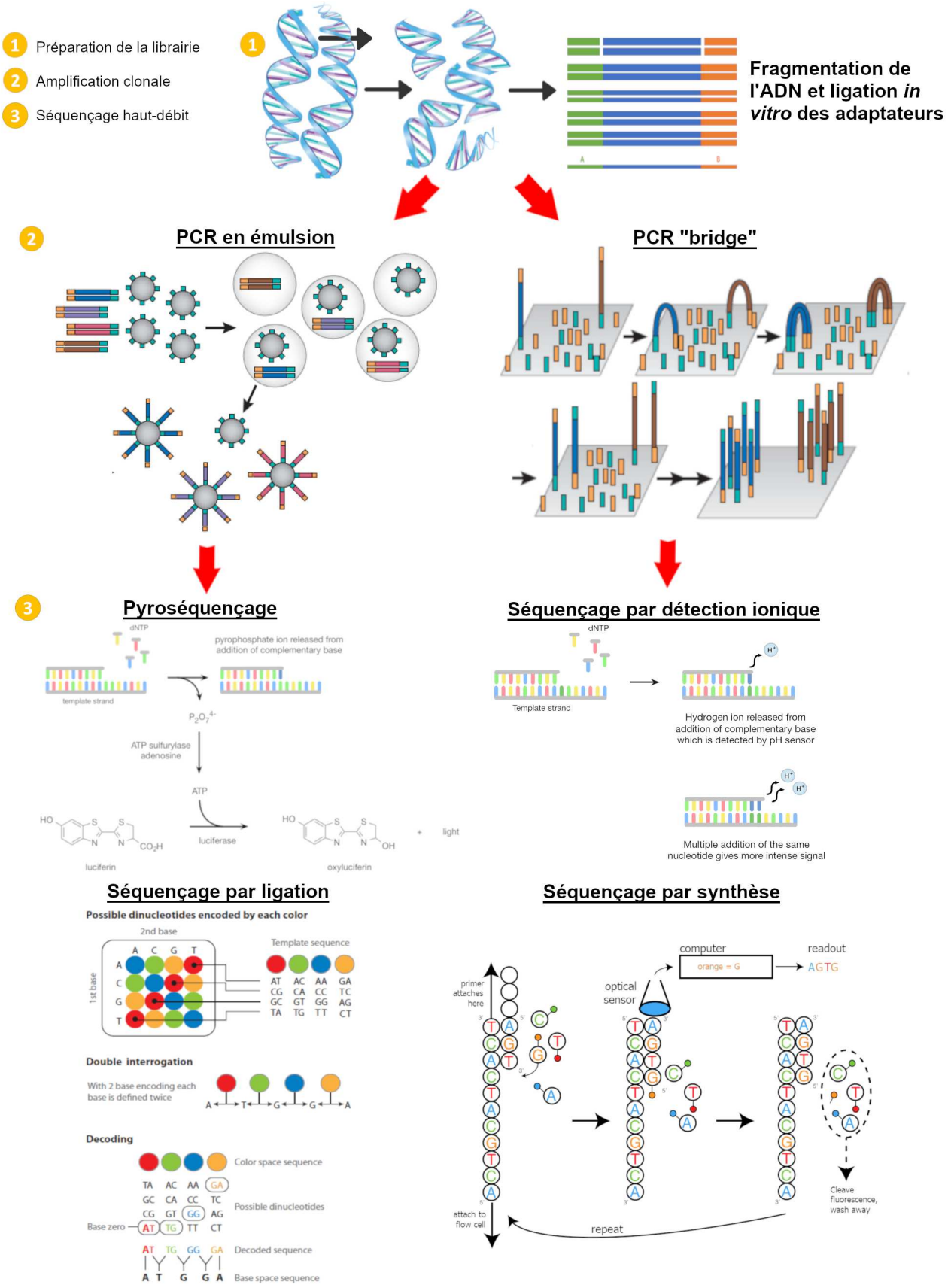

FIGURE 12 – Résumé du séquençage haut-débit pour les 4 méthodes principales de NGS.

2. L'amplification des séquences : la librairie est amplifiée par des méthodes de PCR (Polymerase Chain Reaction) en émulsion (chaque fragment d'ADN double brin est accroché à une bille différente) ou de PCR "Bridge" (chaque fragment d'ADN simple brin est fixé aléatoirement sur une même plaque).

3. Le séquençage : chaque fragment amplifié va alors être lu en fonction de la technologie choisie. Les séquences de fragments lus (appelées les *reads*) sont alors enregistrées avec leur score de qualité (score Phred) dans des fichiers Fastq.

TGS : la 3ème génération de séquençage

De nouvelles technologies de séquençage sont apparues depuis, les TGS (Third Generation Sequencing). Les TGS se basent sur l'inspection directe d'une seule molécule d'ADN. Ils permettent ainsi le séquençage de fragments plus grands que les séquençages NGS, de 5 000 à 15 000 paires de bases en moyenne contre 100 à 800 paires de bases (Lee *et al.*, 2016). Ils existent plusieurs méthodes différentes de séquençage TGS détaillés dans la Figure 13. En revanche, ces techniques produisent encore beaucoup d'erreur de séquençage par rapport au NGS.

Les différentes caractéristiques des trois générations de séquençage sont résumées dans la Table 4.

B.1.2. Génomique et transcriptomique

Les séquençages haut-débit ont permis de nouvelles analyses du génome et du transcriptome (Lee *et al.*, 2013). Au niveau du génome, il est possible de faire du séquençage *de novo* ou d'aligner les séquences sur un génome de référence en utilisant des outils d'alignement comme TopHat2 (Kim *et al.*, 2013), Bowtie2 (Langmead et Salzberg, 2012) ou BWA (Li et Durbin, 2009). Au niveau du transcriptome, les NGS permettent de faire du RNA-Seq (Séquençage d'ARN) en utilisant l'ADN complémentaire (ADNc). On peut au préalable sélectionner des ARN spécifiques comme les ARNm ou les microARN. Comme pour l'ADN, on peut faire un séquençage *de novo* ou aligner les lectures du transcriptome sur un génome de référence ou un transcriptome de référence. Actuellement, pour l'alignement des reads du transcriptome sur le génome, le guide de GATK recommande d'utiliser STAR (Dobin *et al.*, 2013) qui est un aligneur spécialement conçu pour le mapping d'ARN car il fait notamment la prédiction des sites donneurs et accepteurs d'épissage. Les alignements sont alors enregistrés dans des fichiers bam ou cram.

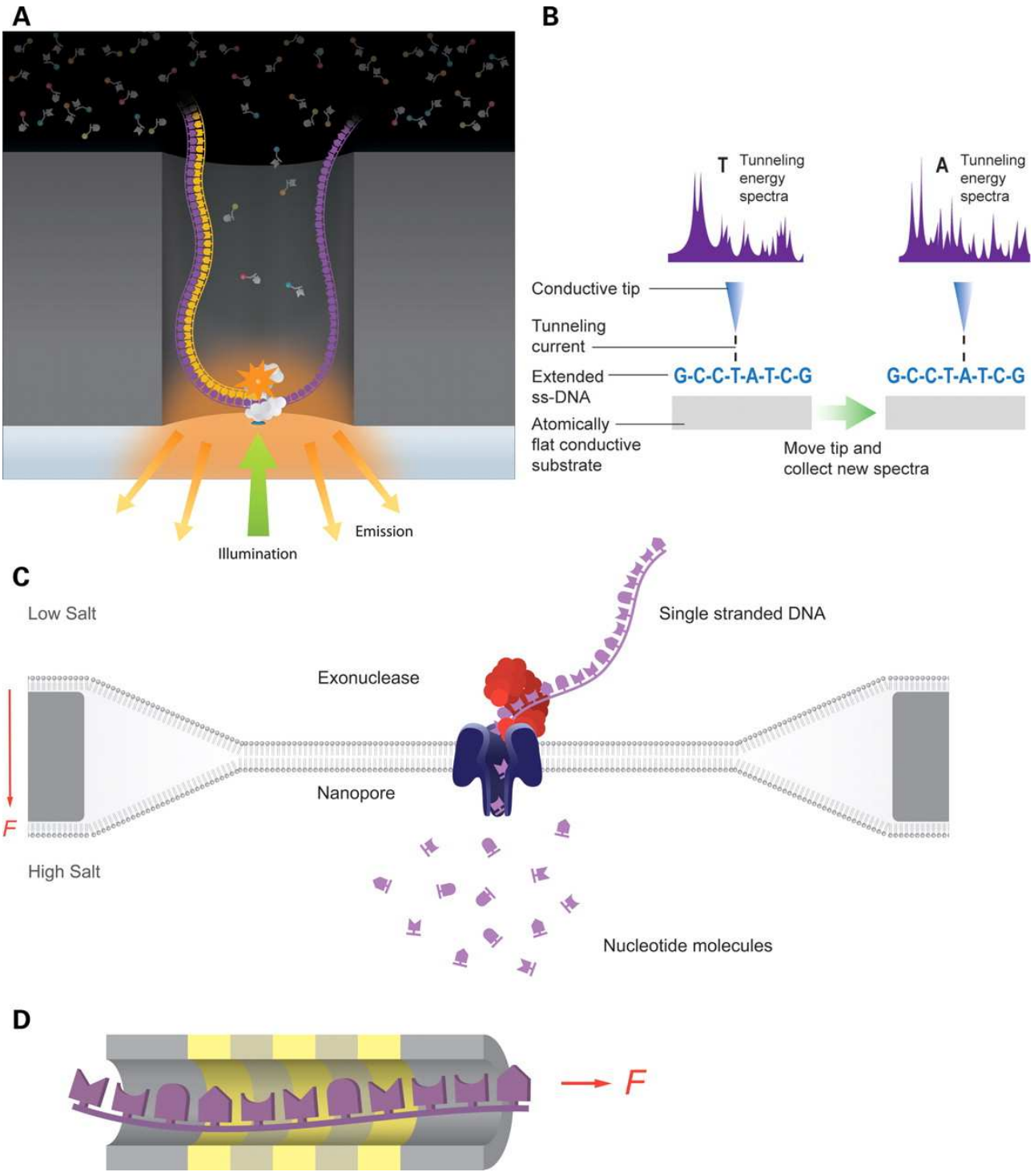

FIGURE 13 – Fonctionnement des différentes technologies TGS par Schadt *et al.* (2010). **A** Méthode PacBio. **B** Utilisation d'un microscope effet tunnel. **C** Méthode Oxford Nanopore. **D** Technologie à transistor ADN d'IBM.

Génération	Première (Sanger)	Deuxième (NGS)	Troisième (TGS)
Résolution	Moyennes de plusieurs copies d'ADN	Moyennes de plusieurs copies d'ADN	Une seule molécule
Exactitude du read brut	Forte	Forte	Modérée
Longueur des reads	Modérée (800–1000 bp)	Courte (100 - 500bp)	Longue (+ de 1000bp)
Débit actuel	Bas	Haut	Modérée
Coût actuel par base	Coût élevé	Coût modique	Coût modique à modéré
Coût actuel par run	Coût modique	Coût élevé	Coût modique
Méthode de RNA-Seq	Séquençage ADNc	Séquençage ADNc	Séquençage direct de l'ARN et séquençage ADNc
Temps de séquençage	Heures	Jours	Heures
Préparation des échantillons	Modérément complexe, ne nécessite pas d'amplification par PCR	Complexe, nécessite une amplification par PCR	De complexe à très simple en fonction de la technologie
Analyse des données	Routine	Complexe à cause d'un gros volume de données et un assemblage compliqué	Complexe à cause d'un gros volume de données et de nouvelles informations à gérer
Résultats principaux	Bases et leur score de qualité	Bases et leur score de qualité	Bases et leur score de qualité, et éventuellement d'autres informations comme la cinétique

TABLE 4 – Comparaison des différentes générations de séquençage traduit de Schadt *et al.* (2010).

B.1.3. Point sur le génome bovin : assemblage

Le tout premier assemblage de référence du génome bovin (*Bos taurus*) Btau_1.0 a été publié en 2004 par le HGCS (Human Genome Sequencing Center) de l'Université Baylor de Médecine (Everts-van der Wind *et al.*, 2004) à partir d'une vache de race Hereford. Le génome bovin est constitué de 29 paires d'autosomes et d'une paire de gonosomes XY. Le tout dernier assemblage a été publié en avril 2018 (ARS-UCD1.2) par l'USDA-ARS (Service de Recherche en Agriculture du Département d'Agriculture des Etats-Unis). Durant ma thèse, j'ai utilisé un assemblage précédent présent sur la base de données Ensembl : UMD3.1 (Zimin *et al.*, 2009). L'annotation du génome de référence UMD3.1 indique que les bovins possèdent 19 994 gènes codants et 3 825 non codants qui produisent en tout 26 740 transcrits différents. Le génome bovin partage 80% de ces gènes avec le génome humain et possède environ 3 milliards de paires de bases (The Bovine Genome Sequencing and Analysis Consortium *et al.*, 2009).

B.2. Importance des polymorphismes : détection des variants

Les polymorphismes sont des variabilités du génome. Il s'agît de l'occurrence de deux à plusieurs formes d'une partie du génome dans la population d'une espèce qui peuvent résulter en des phénotypes différents. Le polymorphisme génétique désigne la coexistence de plusieurs allèles pour un gène ou un locus donné. D'un point de vue strict on différencie un polymorphisme d'un variant ou mutation en fonction de la fréquence dans la population : si un variant est présent chez plus d'un pourcent de la population c'est un polymorphisme. Il existe 3 types principaux de polymorphismes : le polymorphisme nucléotidique (SNP pour Single Nucleotide Polymorphism), l'insertions-délétion de bases (indel) et la variabilité du nombre de copies (CNV pour Copy Number Variation).

B.2.1. Les polymorphismes nucléotidiques (SNPs)

Le polymorphisme nucléotidique est un polymorphisme affectant une seule paire de bases de l'ADN, entre des individus d'une même espèce. Par exemple, à une position précise du génome humain, le nucléotide A peut apparaître chez 80% des individus et le nucléotide G chez 40% des individus, G et A sont des allèles de cette position. Quand un SNP a seulement deux allèles possibles, il s'agît d'un SNP biallélique ; plus rarement un SNP peut avoir 3 allèles possibles il est dit trialléque (Casci, 2010). Un indivu avec ce SNP biallélique peut être hétérozygote, c'est à dire qu'un des allèles est présent sur un chromosome et que l'autre allèle est présent sur le chromosome homologue. Dans le cas de l'exemple précédent, on a 20% d'individus hétérozygotes A/G et donc 80% d'individus homozygotes (60% A/A et 20% G/G). Les SNPs sont un phénomène fréquent chez les mammifères ; chez l'homme, 136 millions de SNPs sont validés sur dbSNP (une base de données regroupant les SNPs détectés - Sherry *et al.*, 2001) ; chez le bovin (*Bos taurus*) 104 millions de SNPs ont été soumis sur dbSNP mais seulement 12 millions ont été validés (version 150 datant de février 2017).

La présence d'un SNP peut influencer l'expression d'un gène de différente façon. Un SNP dans la région promotrice peut influencer l'expression des gènes en affectant les sites de fixation à un facteur de transcription (TFBS - Guo et Jamison, 2005). De manière analogue, un SNP dans la région 3'UTR d'un gène peut affecter la fixation d'un microARN (Saunders *et al.*, 2007). Dans les deux cas, le SNP peut créer un nouveau site de fixation, en supprimer un ou augmenter/diminuer l'affinité du site. Dans la région codante, un SNP peut ne pas affecter la séquence de la protéine (voir Table 2, page 15) il est dit synonyme alors qu'un SNP non-synonyme va modifier la séquence de la protéine soit en changeant l'acide aminé (faux-sens), soit en terminant la séquence par un codon stop (non-sens). Un SNP peut aussi affecter un site d'épissage entraînant généralement un épissage

alternatif (Ju *et al.*, 2015) et il peut aussi affecter la séquence de certains microARN (Saunders *et al.*, 2007) et lncRNA (Miao *et al.*, 2018) quand il est présent dans une région intergénique ou intronique.

B.2.2. Les insertions et délétions (indels)

L'insertion consiste en l'ajout de bases supplémentaires en un endroit du génome et la délétion représente la disparition de bases en un endroit du génome. Comme pour les SNPs, un indel hétérozygote désigne la présence des deux allèles différents sur les chromosomes homologues. Les indel mesurent entre 1 et 10 000 nucléotides (Mullaney *et al.*, 2010), quand ils n'excèdent pas 50 bases, on parle d'un microindel (Gonzalez *et al.*, 2007). Dans les régions codantes, les indels peuvent provoquer des décalages du cadre de lecture entrainant la formation de codon stop prématuré. Toutefois, ils sont plus fréquents dans les régions non codantes. Tout comme les SNPs, les indels peuvent affecter les régions régulatrices, notamment les TFBS (Ribeiro-dos Santos *et al.*, 2015).

B.2.3. Les variabilités du nombre de copie (CNV)

Les variabilités du nombre de copies (ou CNV) désignent un phénomène où un segment est répété de manière variable entre des individus de la même espèce (Sebat *et al.*, 2004). La perte ou le gain de segments varie de 50 paires de bases à plusieurs millions de bases. Des 3 types de polymorphismes présentés dans ce chapitre, les CNV sont les moins fréquents mais ils affectent aussi l'expression des gènes (Bickhart *et al.*, 2012) et leur impact est plus important car il recouvre un plus grand pourcentage du génome. Chez les mammifères, ils assurent une fonction importante dans la variabilité génétique mais ils peuvent être tout autant responasbles de phénotypes liés à des maladies (Ionita-Laza *et al.*, 2009). Des études sur le bovin montrent le rôle des CNV dans la variabilité génétique et suggèrent que certains CNV pourraient être associés à des différences spécifiques à chaque race (Letaief *et al.*, 2017; Bickhart *et al.*, 2012).

B.2.4. Détection des polymorphismes

La détection des variants est accomplie à partir des fichiers d'alignement de séquences et consiste à identifier les différences entre les reads alignés et le génome de référence sur lequel il est aligné. On distingue deux types de détection, la détection des variants de lignée germinale (Bansal, 2010) qui consiste à utiliser le génome de référence standard de l'espèce d'intérêt et la détection des variants somatiques (Roth *et al.*, 2012; Xu, 2018) qui consiste à utiliser comme génome de référence la séquence d'un tissu provenant du même individu. La première méthode permet d'identifier des

variants au sein d'une population et l'hétérozygotie des individus diploïdes alors que la deuxième méthode permet de mettre en évidence le mosaïcisme entre les cellules et de détecter des variants propres au cancer (Cibulskis *et al.*, 2013). Les variants ainsi détectés sont enregistrés dans un fichier GVF (genome variation format), VCF ou BCF (variant/binary call format) avec les informations de génotypage, la position et des informations sur la qualité.

Il existe de très nombreux outils pour la détection de SNPs et d'indels (Sandmann *et al.*, 2017; Xu, 2018). Un des plus utilisés est l'outil HaplotypeCaller de GATK (Genome Analyse Toolkit - DePristo *et al.*, 2011) qui permet d'effectuer une détection des SNPs et des indels somatique ou de lignée germinale. Les étapes principales de la détection des petits polymorphismes (SNPs et indels) sont détaillés dans la Figure 14. La détection concomitante des SNPs et des indels, permet à l'outil de retirer les SNPs présents dans des indels afin de réduire le nombre de faux positifs. Pour réduire encore le nombre de faux positifs, il est recommandé de filtrer les petits variants présents dans des régions contenant des CNVs ; une détection spécifique des SNPs de manière robuste nécessite donc de détecter en parallèle les trois types de polymorphismes présentés dans cette section. Enfin, on peut annoter ces variants pour déterminer les potentielles conséquences qu'ils peuvent avoir sur l'expression des gènes grâce à l'outil VEP (Variant Effect Predictor - McLaren *et al.*, 2016) par exemple. Les différents effets pouvant être prédits par VEP sont détaillés dans la Figure 15.

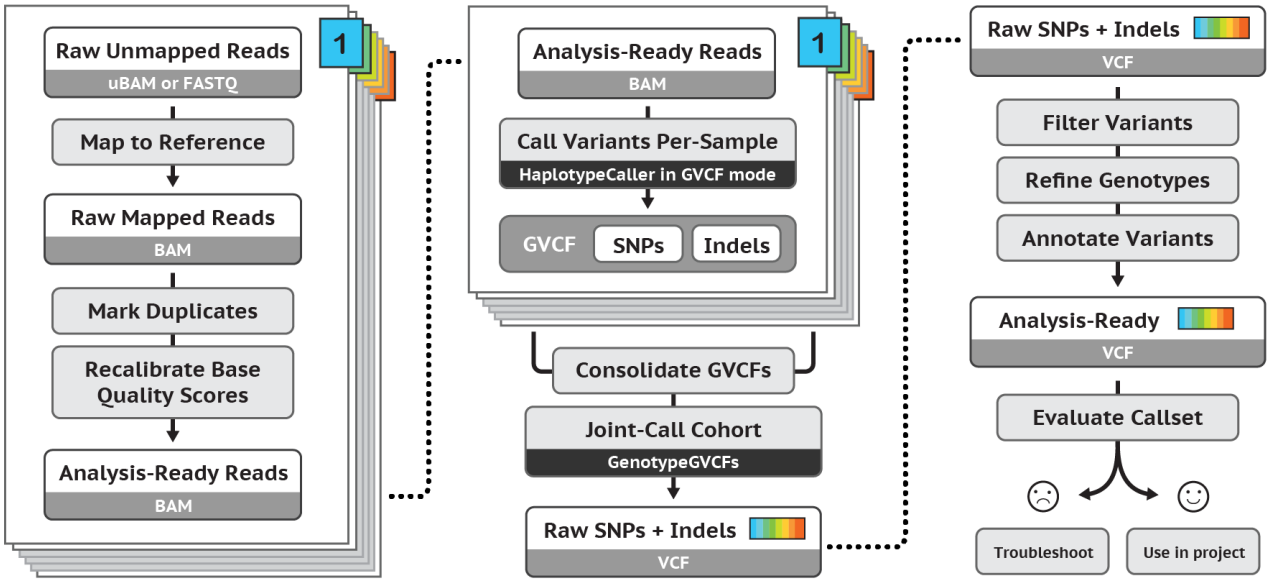

FIGURE 14 – Protocole d'analyse pour la découverte des petits variants (SNPs et indels) de lignée germinale avec les outils de GATK d'après le best practices workflow.

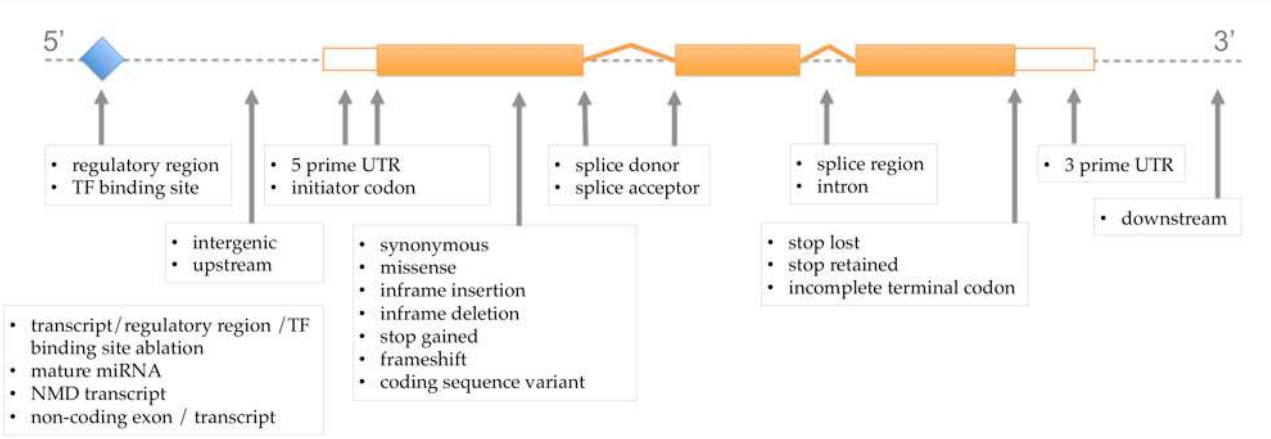

FIGURE 15 – Conséquences prédites par VEP (McLaren *et al.*, 2016).

B.3. Etude bio-informatique de la régulation de l'expression des gènes

A l'aide des données RNA-seq, il existe différentes méthodes pour étudier la régulation de l'expression des gènes. A partir des alignements, il est possible de faire des analyses de l'expression différentielle pour détecter une expression tissu-spécifique ou une différence intra-tissulaire. Avec les variants détectés, il est possible d'analyser des régulations en *cis* avec l'expression allèle-spécifique ou en *trans* avec des analyses eQTL (expression quantitative trait loci). Chez l'homme, il existe un projet à grande échelle ayant pour vocation d'étudier la régulation et l'expression tissu-spécifique : le projet GTEx (Genotype-Tissue Expression - Lonsdale *et al.*, 2013; The GTEx Consortium, 2015). Ce projet met en place une grande base de données avec une grande banque de tissus associées et regroupe un grand nombre d'analyses eQTLs et ASE.

B.3.1. Expression tissu-spécifique : analyse de l'expression différentielle

L'ADN est présent dans presque toutes les cellules de manière quasi identique à quelques mutations près. Toutefois, les gènes exprimés diffèrent d'une cellule à l'autre en fonction du tissu, de l'environnement (infection, stress, alimentation,*etc.*) ou de l'activité de la cellule (Patient, 1990). L'expression tissu-spécifique dépend principalement de la différenciation cellulaire au stade embryonnaire et de modifications épigénétiques qui vont éteindre certains gènes (voir chapitre A.9.2, page 26). Dans les cellules cancéreuses ou infectées, des gènes exprimés vont différer de ceux exprimés dans les cellules saines du même tissu (Zhang *et al.*, 1997; Krishnan et Zeichner, 2004). Il est alors important d'analyser l'expression différentielle afin de déterminer les gènes impliqués dans une maladie, propre à certains tissus ou à différents états cellulaires. Ces analyses au niveau des polymorphismes permettent de déterminer des biomarqueurs liés à une maladie ou à un groupe de

tissus.

A partir des données RNA-seq, on peut effectuer des analyses d'expression différentielle. Une fois les alignements effectués sur le génome, il faut déterminer le niveau d'expression de chacun des gènes pour chacun des tissus ou types de cellules (infectées ou saines). L'outil le plus utilisé pour mesurer le niveau d'expression est HTSeq-count (Anders *et al.*, 2015) car il permet d'estimer le niveau d'expression des gènes, des transcrits ou des exons (Anders *et al.*, 2012) en fonction de l'annotation fournie. Une fois, le comptage de l'abondance pour chaque gène et chaque échantillon obtenu, il faut normaliser les résultats notamment pour supprimer les biais liés à la taille des gènes ou le taux de GC (Hansen *et al.*, 2012). L'utilisation d'une loi binomiale négative permet d'identifier les gènes qui sont différentiellement exprimés entre eux de manière significative. L'outil DESeq2 (Love *et al.*, 2014) permet d'effectuer la normalisation puis les tests statistiques pour estimer le niveau d'expression différentielle. Les résultats peuvent ensuite être représentés à l'aide d'une heatmap couplée à un regroupement hiérarchique afin de déterminer quels groupes de gènes sont plus ou moins exprimés dans un tissu ou un autre.

B.3.2. Expression allèle-spécifiques (ASE)

Dans le génome, certains gènes sont hétérozygotes c'est-à-dire qu'il possède deux allèles différents de ce gène pour chacun des chromosomes homologues. Dans le cas des individus diploïdes comme les mammifères, ces deux allèles proviennent respectivement des deux parents. Au niveau de l'ADN, le ratio de ces deux allèles est généralement de 50 :50, par contre il peut y avoir un déséquilibre dans l'expression des allèles. On parle alors de déséquilibre allélique ou d'expression allèlespécifique (Allele Specific Expression - ASE). L'allèle surexprimé est dit dominant et le sous-exprimé est dit récessif. Le cas extrême de l'expression allèle-spécifique est l'expression monoallélique, lorsqu'un des allèles est complètement inactivé (résultant d'une empreinte parentale par exemple). En dehors de l'empreinte parentale, il peut s'agir de l'effet d'une régulation génétique ayant lieu en *cis*, par conséquent l'étude de l'expression allèle-spécifique est une approche robuste pour détecter l'effet d'une influence génétique agissant en *cis* sur l'expression du gène (Crowley *et al.*, 2015). Plus rarement, un déséquilibre allélique peut également être dû à l'apparition prématuré d'un codon stop provoqué par un variant non-sens (MacArthur *et al.*, 2012).

De nombreux outils de détection ont été mis en place (Gu et Wang, 2015). Voici les principales étapes pour les détecter. Tout d'abord, à partir des variants bialléliques détectés, il faut compter le nombre spécifique de lectures alignées sur chacun des allèles. L'outil ASEReadCounter inclus dans GATK (Castel *et al.*, 2015) permet d'obtenir un comptage précis à partir des fichiers VCF et BAM.

Pour être sûr de bien détecter les régulations en *cis*, il est important de conserver seulement les variants qui sont hétérozygotes pour les données ARN et ADN afin d'éliminer des variants soumis à l'empreinte parentale ou à l'édition d'ARN. Il est nécessaire aussi d'éliminer les biais d'alignements soit en alignant sur le génome parental soit en produisant un génome de référence personnalisé pour chaque individu soit en remplaçant les SNPs dans le génome de référence par un "N" (N-masking) ou une troisième base (Degner *et al.*, 2009; Satya *et al.*, 2012; Wood *et al.*, 2015). Une dernière recommandation est de ne conserver que les variants avec un nombre de reads supérieurs à 10 pour supprimer d'éventuels biais d'alignement dus à une faible expression.

Une fois tous ces filtres appliqués, il faut s'assurer que la différence d'expression entre les deux allèles est significative. Plusieurs tests statistiques existent pour s'en assurer. Le plus simple reste le test binomial sur les variants hétérozygotes bialléliques avec une hypothèse H1 d'un taux de succès différent de $1/2$ (Castel *et al.*, 2015). Sur le même principe, un test du χ^2 peut permettre de détecter un déséquilibre allélique (Li *et al.*, 2012). En possédant les données d'ADN de l'individu testé, une table de contingence 2 x 2 et un test du χ^2 permet de prendre en compte un éventuel biais au niveau génomique (Pastinen, 2010) mais cette méthode nécessite une bonne couverture des lectures pour les échantillons d'ADN. L'idéal serait de connaître pour chaque variant quel est l'allèle paternel et l'allèle maternel. Dans ce cas-là, il est recommandé d'utiliser un test de Student homoscédastique comparant la moyenne d'un premier groupe qui est le ratio des allèles provenant des données RNA-Seq avec la moyenne d'un deuxième groupe qui est la moyenne des allèles provenant des données ADN (Springer et Stupar, 2007). Cette dernière méthode est très pratique pour mettre en évidence l'influence de chaque parent dans le cas de races hybrides (Springer et Stupar, 2007).

De nombreuses études d'ASE ont été effectuées chez l'humain, notamment dans le cadre du projet GTEx (Serre *et al.*, 2008; Chen *et al.*, 2016) mais aussi chez la souris (Lagarrigue *et al.*, 2013; Crowley *et al.*, 2015) ou la mouche (Fear *et al.*, 2016). Des études ASE ont aussi été réalisées chez les animaux d'élevage, notamment chez le porc (Maroilley *et al.*, 2017), la poule (Zhuo *et al.*, 2017) ou le mouton (Ghazanfar *et al.*, 2017). Chez le bovin, Chamberlain *et al.* (2015) ont publié une cartographie des ASE dans 18 tissus différents provenant d'une seule vache de race Holstein.

B.3.3. Analyses eQTL : eGWAS

Un locus de caractères quantitatifs associé à l'expression (eQTL pour expression quantitative trait locus) est une région d'ADN intégrant quelques polymorphismes de petite taille (SNP, microindel) qui affectent le niveau d'expression d'un gène situé en *cis* ou en *trans* (Williams *et al.*, 2007). Ils peuvent être identifiés par des études d'association pangénomique de l'expression (eGWAS pour expression

genome-wide association studies), une méthode d'analyse qui consiste à calculer la probabilité qu'un polymorphisme affecte l'expression des gènes. Toutefois, ce type d'analyse nécessite un grand nombre d'échantillons pour minimiser le taux de faux positifs (Haley et De Koning, 2006).

L'analyse eGWAS se base sur des analyses statistiques de l'expression de chaque gène comme un phénotype quantitatif indépendant. Il est important de tenir compte de la structure familiale et de corriger pour les tests multiples (Kendziorski et Wang, 2006). La puissance statistique des études dépend fortement de la taille de l'échantillon, en effet multiplier la taille de l'échantillon par deux peut augmenter par quatre la puissance de détection des eQTL (Williams *et al.*, 2007).

Les eGWAS permettent la détection des effets en *cis* et en *trans* désignée respectivement sous le terme *cis*-eQTL et *trans*-eQTL. Lors du premier chapitre, on a vu que les causes de régulation sont nombreuses, elles sont résumées dans la Figure 16. Toutefois, il est parfois difficile de déterminer si l'effet est en *cis* ou en *trans* avec des eGWAS et la limite entre les deux est fréquemment déterminée par une distance choisie arbitrairement. Bien que les variants *trans* soient généralement plus présents sur les autres chromosomes (Pai *et al.*, 2015) certains peuvent être très proches du gène affecté. Inversement, certains variants en *cis* peuvent être assez éloignés du gène affecté jusqu'à plusieurs milliers de kilobases (Smith *et al.*, 2013). Pour lever cette ambiguïté, Rockman et Kruglyak (2006) recommandent de distinguer local/distant au lieu de *cis*/*trans*.

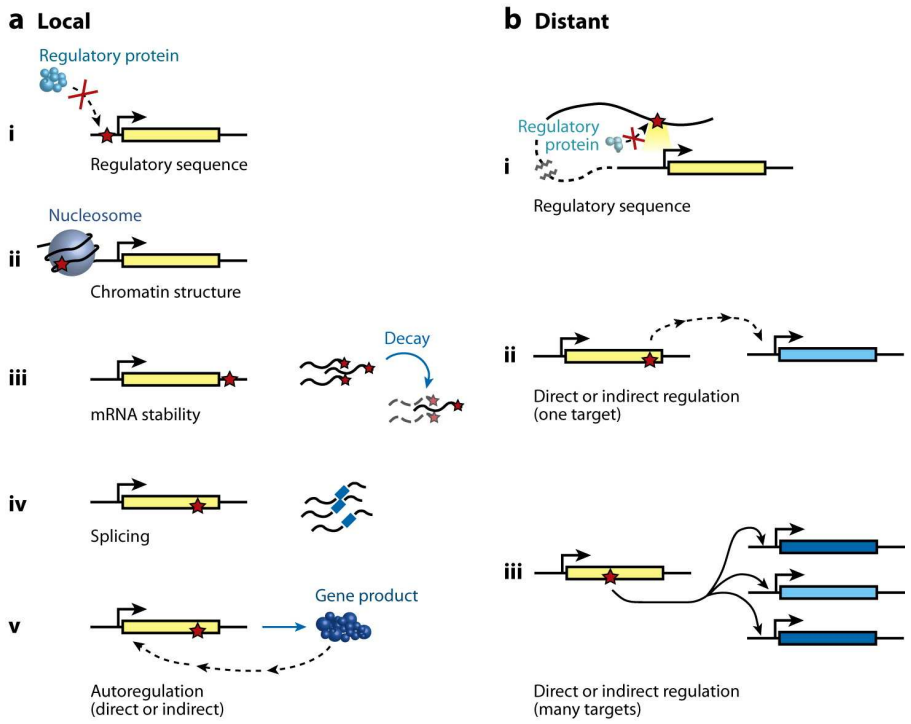

FIGURE 16 – Mécanismes des variations régulatrices en local (cis) et en distant trans par (Skelly *et al.*, 2009).

La détection des *cis*-eQTL peut être complétée par les analyses ASE en caractérisant directement

les variants agissant en *cis*. Ces analyses sont en effet plus puissantes grâce à un meilleur contrôle intra-individuel qui élimine ainsi les influences génétique en *trans* et les facteurs environnementaux (Pastinen, 2010).

De nombreuses eGWAS ont été réalisées chez l'humain, notamment avec le programme GTEx (Zou *et al.*, 2012; Sabbagh *et al.*, 2016; Grigoryev *et al.*, 2015). Chez les mammifères, on trouve des études sur la souris (Schadt *et al.*, 2003) ou le porc (Ponsuksili *et al.*, 2014; Maroilley *et al.*, 2017), en revanche, très peu d'études ont été effectuées chez le bovin. Jusqu'à présent, seulement deux ont été publiées dont une très récemment : Lopdell *et al.* (2017) sur des vaches laitières de race Frisonne, Jersiaise et mixte (Jersiaise x Frisonne) pour étudier les SNPs liés à la composition du lait en lactose et Higgins *et al.* (2018) sur des bovins irlandais (principalement des hybrides de race Charolaise, Limousine, Bleu-Belge, Simmental et Angus) pour étudier les SNPs liés à l'efficacité alimentaire.

B.4. Validation expérimentales de la régulation de l'expression des gènes

Les différentes analyses *in silico* vu précédemment permettent de détecter un grand nombre de variants impliqués dans l'expression des gènes. Toutefois, bien qu'en constante évolution, un nombre significatif de faux positifs persiste. Il est donc important de valider expérimentalement l'effet de ces variants par des méthodes *in vitro* et/ou *in vivo*.

B.4.1. Gènes rapporteurs : mesure de l'expression d'un gène

Méthode de base

Un gène rapporteur est un gène codant une protéine pouvant être observés aisément de manière expérimentale. Ces gènes doivent être étrangers au génome de l'espèce étudiée, le signal produit doit être quantifiable et le signal produit doit être rapide et précis afin de permettre de suivre en temps réel l'expression du gène et de situer avec précision où il est exprimé. Il existe deux types classiques de gène rapporteur en fonction du signal produit :

- les gènes codant une protéine qui produit un signal lumineux, comme de la fluorescence avec la GFP (Green Fluorescent Protein - Osorio et Bionaz, 2017) ou de la bioluminescence avec la luciférase (Koo *et al.*, 2007) ;
- les gènes codant des enzymes dont l'activité entraîne une coloration de la cellule, comme le gène GUS (Koo *et al.*, 2007) ou lacZ (West *et al.*, 2015).

Les gènes rapporteurs permettent de quantifier l'expression d'un gène ou d'un promoteur en transfectant des cellules. Dans le premier cas, le gène rapporteur est ajouté au gène à étudier formant ainsi un gène de fusion qui sera transcrit en un seul ARNm ; lors de la traduction, le gène rapporteur va permettre de déterminer le taux d'expression du gène ciblé mais il est nécessaire pour cela que la fusion n'influe pas sur les domaines des deux protéines fusionnées (Osorio et Bionaz, 2017). Dans le deuxième cas, le gène rapporteur est placé à la suite du promoteur, permettant de quantifier l'activité du promoteur (Jugder *et al.*, 2016).

Cette méthode possède quelques désavantages : premièrement, suivant le type cellulaire la transfection peut être plus ou moins efficace sur les cellules soumises à cette technique, la fusion créée ne doit pas modifier l'activité de la protéine produite et il est complexe de tester plusieurs gènes en même temps. Pour pallier ce dernier problème, une technique a été mise en place : le MPRA (Massively Parallel Reporter Assay - Melnikov *et al.*, 2012).

Analyse en masse : MPRA

L'approche expérimentale MPRA (Figure 17) est une méthode qui repose sur le clonage en masse de milliers à des centaines de milliers de séquence d'ADN régulatrices candidates (Melnikov *et al.*, 2014). Cette méthode permet d'étudier un grand nombre de séquences allèle-spécifiques correspondant aux rSNPs (SNPs régulateurs) à tester en les intégrant dans un plasmide contenant un gène rapporteur. Pour chaque séquence étudiée avec les rSNPs d'intérêt, on associe une étiquette qui sera transcrite. Après clonage et transfection de la population de plasmides dans des cellules, l'ARN est extrait et du séquençage RNA-Seq réalisé afin de compter ces étiquettes. Après normalisation, il est ensuite possible de déterminer pour chaque rSNP testé, si les différentes séquences allèle-spécifiques testées ont un effet différent. Cette approche permet donc d'analyser plusieurs milliers de rSNPs candidats en parallèle.

B.4.2. Pyroséquençage : génotypage des polymorphismes

La méthode de pyroséquençage (Figure 12, page 30 et Figure 18) permet de quantifier la proportion d'un nucléotide pour chaque position donnée (Ronaghi *et al.*, 1998). En effet, lors du séquençage un capteur CDC (Charge-Coupled Device) va capter le signal lumineux produit par la luciférase et retranscrire ce signal sous forme de pic sur un pyrogramme. La hauteur de ce pic représente l'intensité lumineuse du signal qui est elle-même proportionnelle à la quantité de nucléotide incorporés en même temps. Grâce à cette quantification, le pyroséquençage peut permettre de détecter un

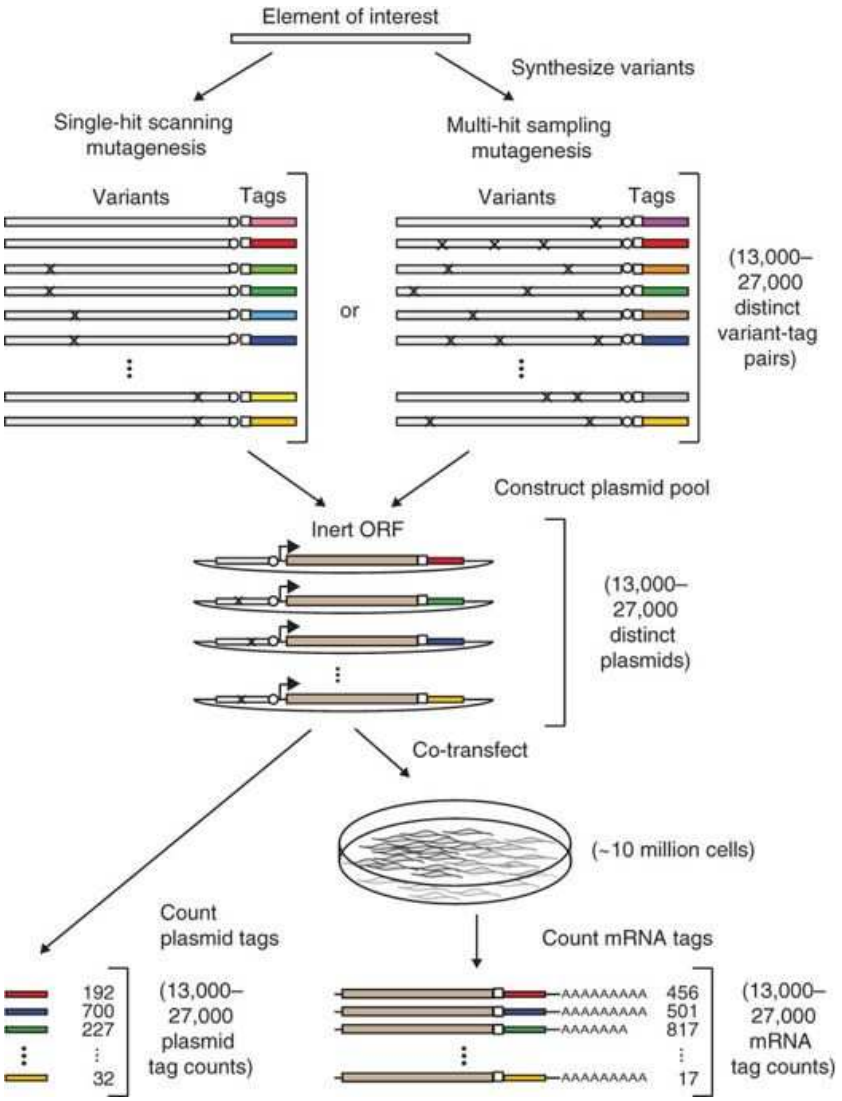

FIGURE 17 – Aperçu de l'approche MPRA selon Melnikov *et al.* (2012).

polymorphisme (Royo et Galán, 2009) ou un gène soumis à ASE ou à l'empreinte parentale (Yang *et al.*, 2013, 2016). Il permet aussi de détecter la méthylation de l'ADN (Delaney *et al.*, 2015).

Pour séquencer un brin spécifique de l'ADN double-brin avec la méthode du pyroséquençage, il est nécessaire qu'une des deux amorces soit étiquetées en 5' avec de la biotine afin d'isoler le bon brin. Les transcrits sont étudiés en utilisant l'ADN complémentaire. Malheureusement, le pyroséquençage est une méthode coûteuse en raison des prix des différentes amorces biotinylées. Royo *et al.* (2007) ont mis en place un protocole de pyroséquençage qui utilise une seule et même amorce biotinylée afin de réduire les coûts et le temps de la technique (Figure 19).

B.4.3. ChIP-Seq : étudier la fixation des facteurs de transcription

Le séquençage ChIP (immunoprécipitation de la chromatine) ou ChIP-Seq est une méthode d'analyse des interactions de protéine avec l'ADN (Barski *et al.*, 2007). Le ChIP-Seq combine im-

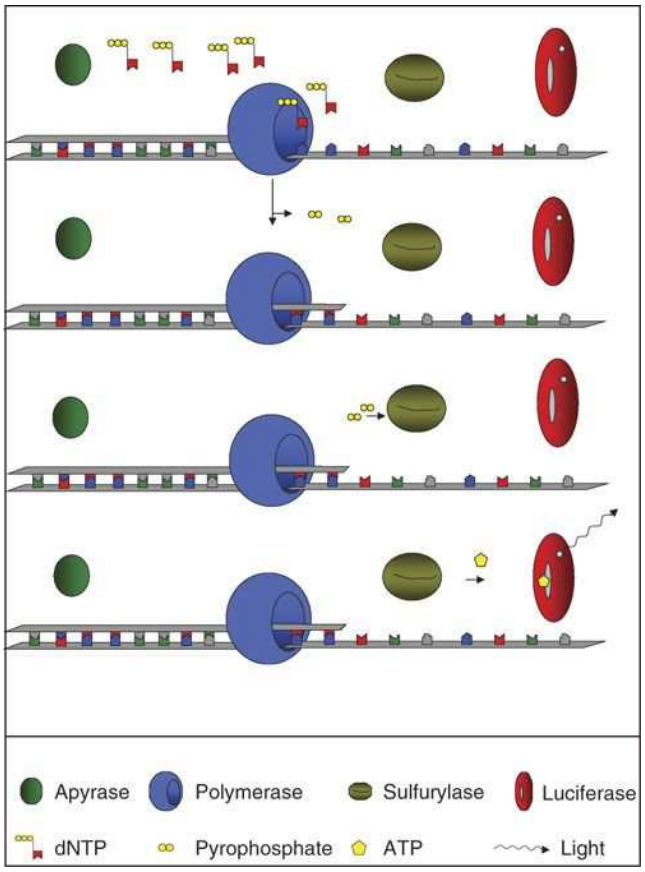

FIGURE 18 – Base moléculaire de la réaction de pyroséquençage par Royo *et al.* (2007)

munoprécipitation de la chromatine avec le séquençage NGS pour identifier les sites de fixation des protéines associées à l'ADN. Cette technologie permet d'aligner les sites de fixation précisément pour n'importe quelle protéine d'intérêt se liant à l'ADN, notamment les facteurs de transcription (Hunt *et al.*, 2014). Elle permet aussi de cartographier l'épigénome en permettant de positionner les nucléosomes et d'identifier les modifications des histones (Park, 2009).

Une fois le ChIP-Seq fait (détaillé en Figure 20), les résultats sont analysés par des méthodes de séquençage. Les lectures sont d'abord alignées sur le génome de référence, puis comptées. A partir du comptage, on détecte des pics dans le génome qui représente les zones enrichies par l'alignement des lecture dû à la fixation de la protéine sur l'ADN. Quand la protéine est un facteur de transcription, la zone enrichie contient le TFBS (Hunt *et al.*, 2014).

B.4.4. La correction de séquence génomique ou Genome Editing

La correction de séquence génomique ou Genome Editing (traduit maladroitement en édition du génome) désigne un ensemble de méthodes visant à manipuler le génome afin d'en réécrire une portion de sa séquence (Blanc *et al.*, 2002). Ces méthodes offrent de nouvelles applications dans l'ingénierie du génome (Hsu *et al.*, 2014) et permettent notamment de tester les fonctions des

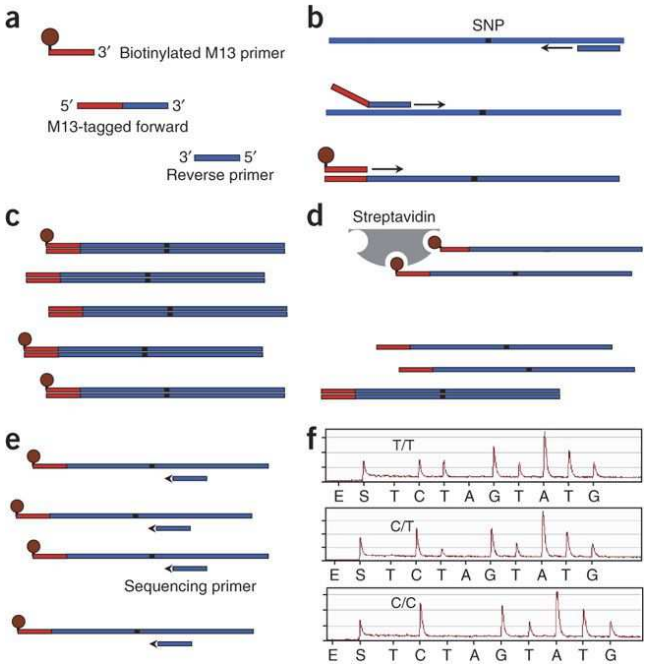

FIGURE 19 – Diagramme du protocole de pyroséquençage utilisant une amorce M13 universel biotinylée par Royo *et al.* (2007). **a**) Les 3 amorces impliquées dans la PCR. **b**) Durant la PCR, l'extrémité M13 est incorporée dans la séquence de l'amplicon. **c**) Le ratio d'amplicons biotinylés ou non dépend de la stœchiométrie de la réaction, de la température et de la concentration saline. d) Les amplicons biotinylés sont capturés par la streptavidine et l'ADN est dénaturé en simple-brin. e) Le brin est relâché et combiné avec l'amorce de pyroséquençage. f) Les résultats sont sous forme de pyrogramme avec le génotype associé : T/T, C/T et C/C

variants régulateurs (Knight, 2014). Elles regroupent les techniques de génie génétique qui consistent à insérer, remplacer ou retirer une à plusieurs séquence d'ADN à un endroit précis du génome en utilisant des nucléases (protéines chargées de couper l'ADN). Les nucléases sont des enzymes pouvant cliver l'ADN au niveau des liaisons entre deux nucléotides. Ces coupures provoquent des cassures double-brin (DSBs pour Double-Strand Breaks) de l'ADN et sont en général couplées avec des mécanismes de réparation de l'ADN afin de modifier le locus ciblé.

Dans le cadre du Genome Editing, deux mécanismes de réparation de l'ADN sont utilisés (Figure 21A) : la jonction d'extrémité non homologue (NHEJ pour Non-Homologous End-Joining - Liang *et al.*, 1998) et la recombinaison homologue (RH ou HDR pour Homology-Direct Repair Renkawitz *et al.*, 2014). La RH est un type de recombinaison génétique où l'ADN est remplacé par des séquences identiques ou similaires et qui est notamment utilisé lors du processus de méiose. Contrairement à la recombinaison homologue, la NHEJ ne restaure pas la séquence initiale d'ADN endommagé mais assure seulement la continuité de celle-ci ce qui peut entraîner l'apparition d'indels.

Il existe 4 familles principales de nucléases artificiellement modifiées développées pour pratiquer le Genome Editing (Esvelt et Wang, 2013) :

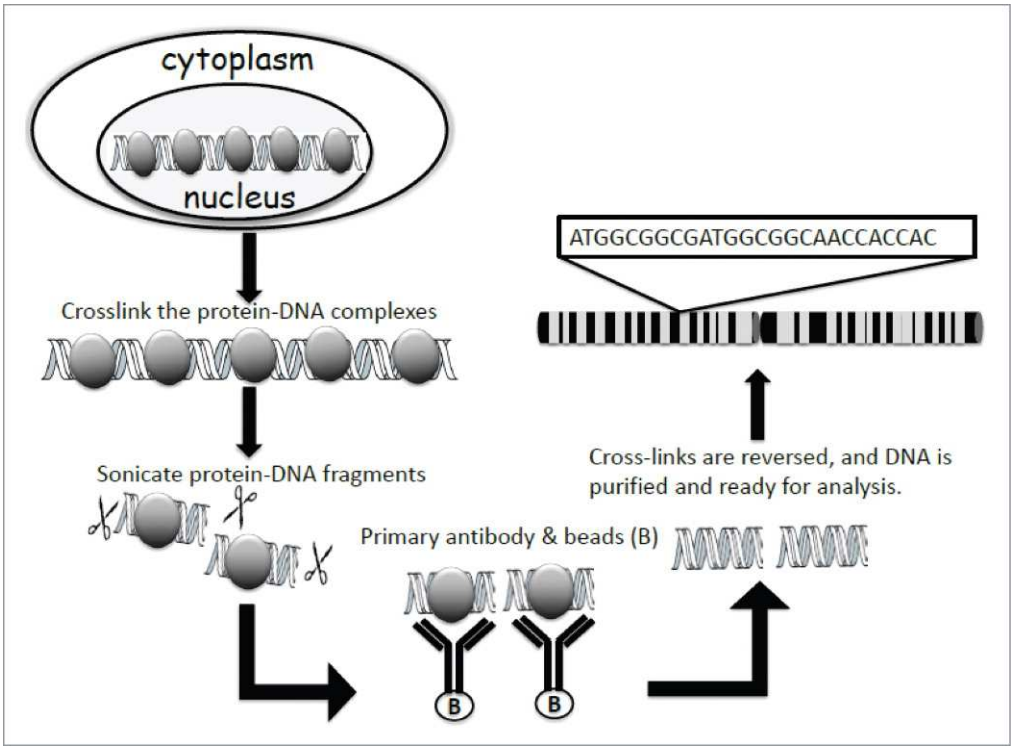

FIGURE 20 – Principe de l'analyse ChIP-Seq par Mundade *et al.* (2014). Le complexe protéine-ADN est réticulé dans le noyau cellulaire afin que la protéine d'intérêt et son site de fixation à la chromatine puissent être fixés. Après la lyse des cellules, le complexe protéine-ADN est découpé par sonication en fragments de 200/1000 paires de bases et ces fragments sont immunoprécipités à l'aide d'anticorps spécifiques. La réticulation du complexe est alors inversée, l'ADN est ensuite purifié puis séquencé avec des méthodes NGS.

1. Les méganucléases (Silva *et al.*, 2011) ;
2. Les nucléases en doigt de zinc (ZFN ou zinc finger nuclease - Figure 21B - Kim *et al.*, 1996) ;
3. Les nucléases effectrices de type activateur de transcription (Transcription activator-like effector nuclease ou TALEN - Figure 21B - Boch *et al.*, 2009) ;
4. Le système CRISPR/Cas9 (CRISPR pour Clustered Regularly Interspaced Short Palindromic Repeats et Cas9 pour CRISPR associated protein 9 - Figure 21C - Jinek *et al.*, 2012).

Grâce aux méthodes de Genome Editing et surtout avec le système CRISPR/Cas9, il est possible d'étudier les mécanismes de régulation d'un polymorphisme *in cellulo* ou *in vivo*. En effet, il est possible de mettre en évidence un SNP impliqué dans une expression allèle-spécifique (Courtney *et al.*, 2016) ou de comprendre le fonctionnement d'un SNP dans la région régulatrice qui agit en *cis* sur l'expression des gènes (Soldner *et al.*, 2016; Knight, 2014).

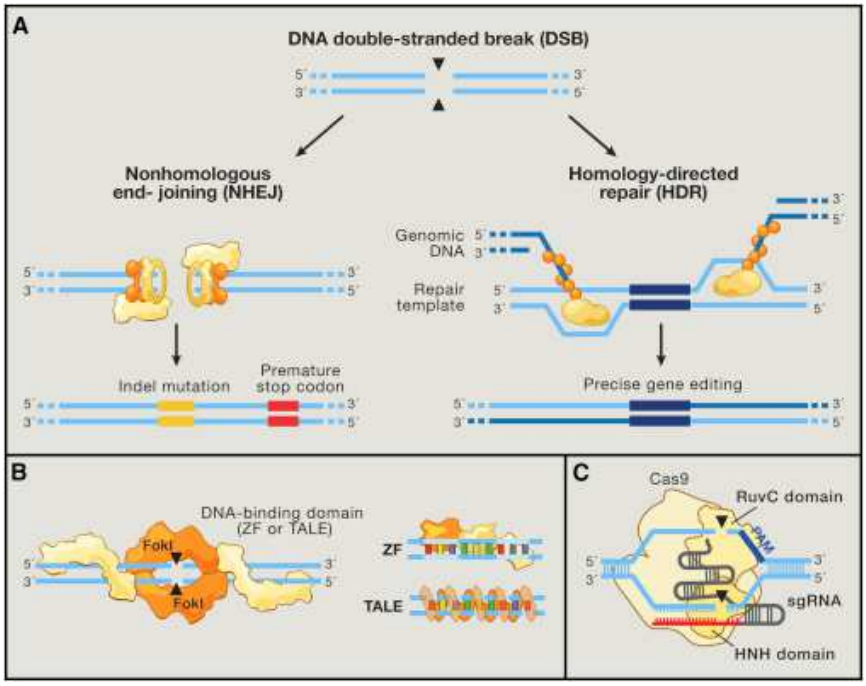

FIGURE 21 – Technologies de Genome Editing exploitant la réparation endogène de l'ADN par Hsu *et al.* (2014). **A**) Les cassures double-brins de l'ADN (DSBs) sont réparées soit par la jonction d'extrémités non homologues (NHEJ) soit par la recombinaison homologue (HDR pour Homology-Direct Repair). **B**) Les protéines en doigt de zinc (ZF pour Zinc Finger) et les effecteurs de type activateur de transcription (TALEs pour Transcription Activator-Like Effectors) présentent naturellement des domaines de liaison à l'ADN qui peuvent être assemblés de façon *ad hoc* pour cibler une séquence précise. Certaines des protéines de liaison à l'ADN peuvent être fusionnées à l'endonucléase FokI pour programmer des nucléases spécifiques à un site. **C**) La nucléase Cas9 est localisée sur des séquences spécifiques d'ADN à l'aide de séquence guide sur son ARN guide (en rouge), directement apparié avec l'ADN ciblé. La liaison d'un motif PAM (protospacer-adjacent motif) en aval du locus ciblé permet de diriger les cassures double brins de l'ADN par l'action de Cas9.

Stratégies de la these

Mon projet de thèse au sein de l'UMR GABI (Génétique Animale et Biologie Intégrative) a pour objectif d'approfondir les connaissances sur la régulation de l'expression des gènes chez le bovin. Ces recherches pourront jouer un rôle dans l'identification de variants génétiques associés à des phénotypes d'intérêt, notamment car la plupart de ces variants se trouve dans la région régulatrice des gènes.

L'objectif global de ce travail est d'identifier chez le bovin et de valider à grande échelle les polymorphhismes qui altèrent potentiellement la régulation de l'expression des gènes et affectent des phénotypes d'intérêt.

Données

Au cours de la thèse, j'ai utilisé deux types de données de séquençage haut-débit provenant de deux races différentes : limousine et holstein.

La race limousine est une race rustique française allaitante, elle est élevée pour la production bouchère et elle est la deuxième race allaitante élevée en France derrière la charolaise. Pour notre étude, j'ai étudié les séquences du génome complet de 19 taurillons limousins ainsi que le séquençage du transcriptome complet du muscle *Longissimus dorsi* qui se situe le long de la colonne vertébrale.

La race holstein est une race laitière. Elle est la race bovine la plus élevée en France ainsi que dans le monde. Pour notre étude, j'ai étudié les séquences du génome complet de 6 vaches holstein ainsi que le séquençage du transcriptome complet de 8 tissus différents : muscle, foie, poumon, rein, rate, cœur, utérus et ovaire, soit 46 échantillons (les données du transcriptome néphrétique étant absentes pour deux individus).

Méthodologie globale

Les étapes définies pour ce travail de thèse sont les suivantes :

1. Je détecte les variants (SNPs et indels).
2. J'identifie les SNPs qui sont soumis à ASE.
3. J'identifie les SNPs dans les régions régulatrices (rSNPs), principalement dans les TFBS ou site de fixation à un microARN.
4. Je calcule le déséquilibre de liaison (LD pour LInkage Desequilibrium) entre SNPs régulateurs (rSNPs) et SNPs dans la région codante (cSNPs).

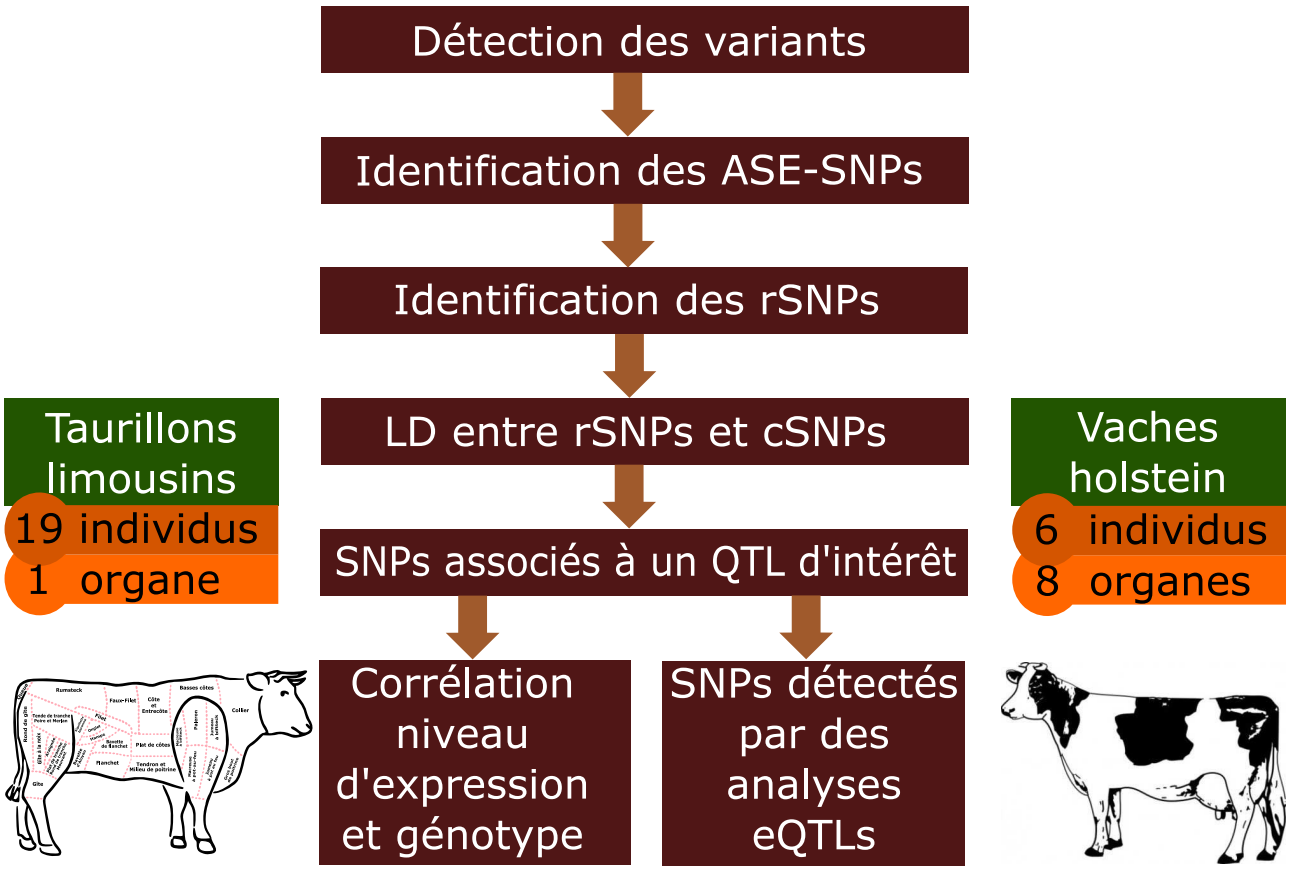

FIGURE 22 – Stratégie globale de la thèse.

5. Je tente de trouver un lien entre des SNPs et des QTL d'intérêt.
6. Je calcule les corrélations entre les niveaux d'expressions et les génotypes d'un SNP pour tous les individus. (limousin seulement).
7. Je compare les ASE-SNPs détectés avec des analyses eQTLs (holstein seulement).

L'ensemble de la stratégie est décrit dans la Figure 22.

Article 1 : Etude de l'expression allèle-spécifique dans le muscle bovin

Article soumis à Scientific Reports le 5 juillet, révision majeure le 17 Octobre.

Objectifs et méthodes

L'objectif de cette étude était d'établir une liste d'ASE-SNPs présent dans le transcriptome musculaire bovin puis d'en relier certains à un SNP régulateur ou à un phénotype d'intérêt associés à la croissance ou à la tendreté de la viande.

Pour cette étude, seules les données provenant des taurillons limousins ont été utilisées. Tout d'abord, j'ai détecté les SNPs montrant une expression allèle-spécifique (ASE-SNPs) à l'aide d'un pipeline en Python utilisant d'autres outils existants (Figure 23). Puis, à partir des ASE-SNPs, on a tenté de caractériser des SNPs régulateurs potentiellement causaux en 3'UTR (dans un site de fixation à un miRNA) et dans la région promotrice (TFBS).

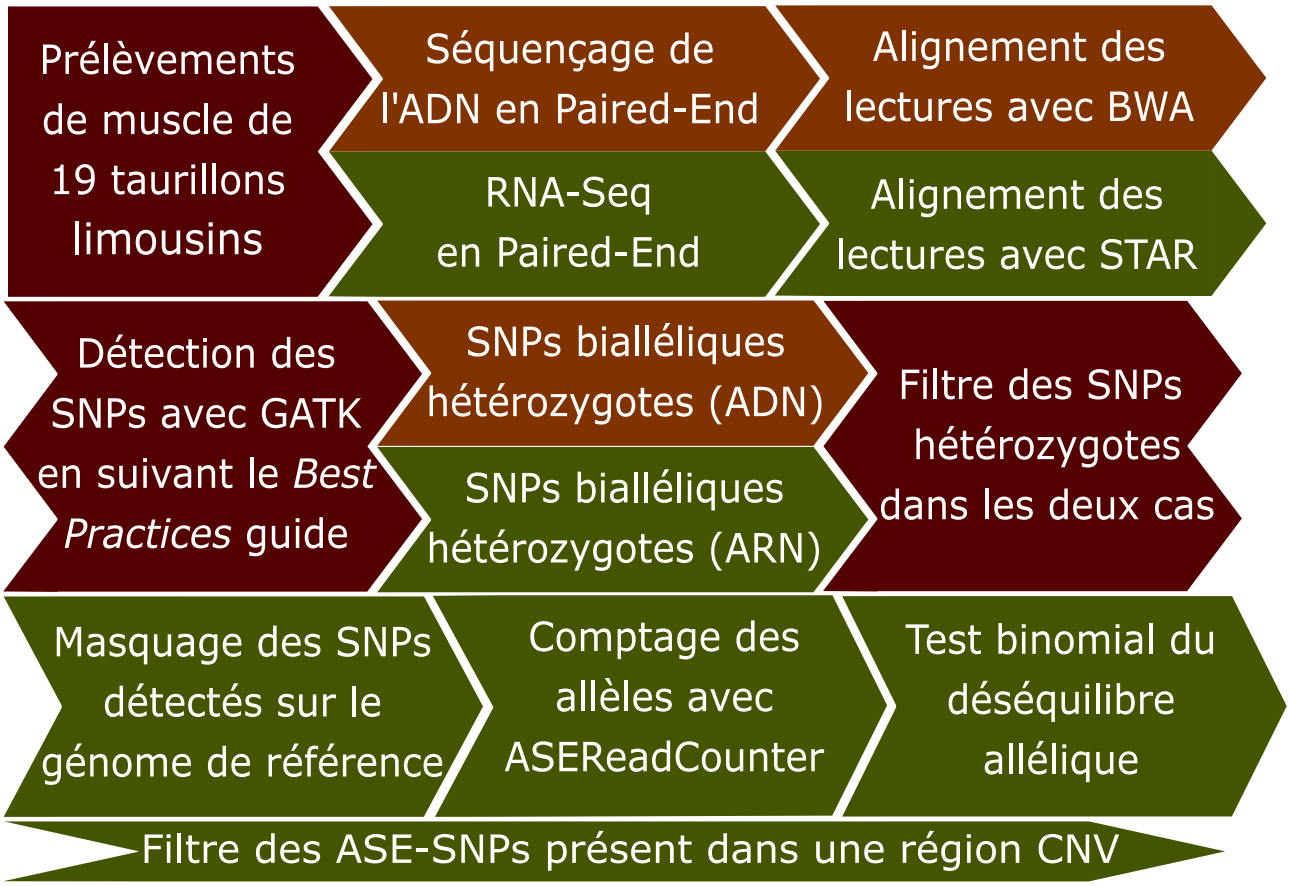

FIGURE 23 – Diagramme du pipeline pour détecter des ASE-SNPs.

Résultats

L'analyse ASE a montré un déséquilibre allélique pour 5 658 SNPs à l'aide des données RNA-Seq et WGS. Ces ASE-SNPs sont présents dans 13% des gènes (2 451) avec une expression détectable dans le muscle et ces gènes avec ASE-SNPs ont principalement des fonctions liées au muscle ou à la structure cellulaire. Cette étude a permis de mettre en évidence une répartition homogène sur le génome bovin et une fréquence non négligeable de ce phénomène. De plus, 37% des ASE-SNPs ont été détectés au sein de régions associées à des traits de carcasse ou de qualités des viandes ainsi que dans des gènes connus comme étant potentiellement impliqués dans le développement musculaire, notamment *AOX1*, *PALLD* et *CAST*. 5 ASE-SNPs sur 6 testés ont été validés par pyroséquençage dont un ASE-SNP dans le gène *PALLD* et deux dans le gène *CAST*. Des approches *in silico* pour détecter des rSNPs ont permis de mettre en évidence 11 SNPs, en déséquilibre de liaison avec un ASE-SNP, dans 6 gènes différents qui modifie l'affinité de fixation de microARN.

Conclusion

Nous avons réalisé une analyse ASE sur le génome complet et le transcriptome musculaire de 19 individus limousins. Cette analyse permet de montrer que le déséquilibre allélique est répandu dans le muscle bovin. Cette étude montre que l'approche ASE peut faciliter l'identification de SNPs candidats régulateurs agissant en *cis*. Toutefois, des travaux supplémentaires sont nécessaires pour valider un plus grand nombre d'ASE-SNPs et pour étudier l'impact des polymorphismes dans les sites de fixation aux microARN ou aux facteurs de transcription.

Survey of allele specific expression in bovine muscle

Gabriel M. Guillocheau[1,*], Abdelmajid El Hou[1], Cédric Meersseman[1,2], Diane Esquerré[3], Emmanuelle Rebours[1], Rabia Letaief[1], Morgane Simao[1], Nicolas Hypolite[1], Emmanuelle Bourneuf[1,4], Nicolas Bruneau[1], Anne Vaiman[1], Christy J. Vander Jagt[5], Amanda J. Chamberlain[5], and Dominique Rocha[1,+]

[1]GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[2]GMA, INRA, Université de Limoges, 87060 Limoges, France

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

[4]CEA, DRF/iRCM/SREIT/LREG, Jouy-en-Josas, France

[5]Agriculture Victoria Research, AgriBiociences Centre, Bundoora, Victoria, Australia

[*]gabriel.guillocheau@inra.fr

[+]dominique.rocha@inra.fr

ABSTRACT

Allelic imbalance is a common phenomenon in mammals that plays an important role in gene regulation. An Allele Specific Expression (ASE) approach can be used to detect variants with a *cis*-regulatory effect on gene expression. In cattle, this type of study has only been done once in Holstein. In our study, we performed a genome-wide analysis of ASE in 19 Limousine muscle samples. We identified 5,658 ASE SNPs (Single Nucleotide Polymorphisms showing allele specific expression) in 13% of genes with detectable expression in the *Longissimus thoraci* muscle. Interestingly, we found allelic imbalance in *AOX1*, *PALLD* and *CAST* genes. We also found 2,107 ASE SNPs located within genomic regions associated with meat or carcass traits. In order to identify causative *cis*-regulatory variants explaining ASE, we searched for SNPs altering binding sites of transcription factors or microRNAs. We identified one SNP in the 3'UTR region of *PRNP* that could be a causal regulatory variant, modifying binding sites of several miRNAs. We showed that ASE is frequent within our muscle samples. Our data could be used to elucidate the molecular mechanisms underlying gene expression imbalance.

Introduction

Gene regulation is a fundamental process in the development and maintenance of organisms. In mammalian genomes the variability of gene expression is a current phenomenon[1,2]. It is therefore important to study this variability in order to understand gene regulation. There are different approaches to such

studies: expression quantitative trait loci (eQTLs) and Allele Specific Expression (ASE) analyses. The combination of both approaches is highly effective at locating *cis*- and *trans*- regulation of gene expression.

An expression quantitative trait locus (eQTL) is a DNA region with some nucleotide sequence differences (Single Nucleotide Polymorphisms, insertion, deletion) that affects the expression level of a gene in *cis* or *trans*. They can be identified by expression genome-wide association studies (eGWAS), an analysis method computing the likelihood of a polymorphism affecting gene expression. Unfortunately this type of analysis needs a large number of samples to minimize false-positives[3]. Many human eQTL mapping studies have been carried out[4–6] including the recent Genotype-Tissue Expression (GTEx) project[7]. However in cattle there is a lack of studies. So far, there has been only one performed in dairy cattle, in Holstein-Friesians (HF), Jerseys (J) and HFxJ crossbreeds[8].

Allele specific expression (allelic expression or allelic imbalance) analysis is a robust approach to quantify expression variation between the two haplotypes of a diploid individual distinguished by heterozygous sites[9]. This approach is complementary to identifying variants affecting gene expression with eQTL studies because we can use a smaller number of samples[10]. Genome-wide studies of ASE have been performed in different species (human[11], mouse[12] or fruit fly[13]) including livestock species (pig[14], chicken[15] or sheep[16]). In addition, some ASE genes were detected to impact economically important traits[10,17].

In cattle, only two studies have been performed so far, both in Holstein. In the first study, they discovered 473 ASE SNPs across 5 bovine blastocysts (among 2,524 different heterozygous SNPs)[18]. In the second study, they detected 19,082 ASE SNPs (1,060 on average per tissue) across 18 different tissues from one lactating Holstein dairy cow[19].

In our study, we performed a genome-wide investigation of ASE using 19 Limousine calf muscle samples. We distinguished between imprinting (parental mono-allelic expression) and allele specific expression (not mono-allelic expression) to focus on the later. We used whole-genome sequences (WGS) and RNA-Seq data from these 19 muscle samples in our analysis. To the best of our knowledge, it is the first ASE survey in a beef breed and the largest number of different animals.

Materials and Methods

Animals and tissue samples

Nineteen Limousine bull calves were selected from a large study on the genetic determinism of beef and meat quality traits[20]. They were fattened in a single feedlot and fed *ad libitum* with wet corn silage. They were humanely slaughtered in an accredited commercial slaughterhouse when they reached 16 months. *Longissimus thoracis* (LT) muscle samples were dissected immediately after death and tissue samples were snap frozen in liquid nitrogen and then stored at $-80\,^{\circ}\mathrm{C}$. The animals used in this study were beef animals raised for commercial reasons from a previous study[20] and were slaughtered by certified slaughterhouses in accordance with French animal protection regulations (Code Rural, Articles R214-64 to R214-71; Legifrance, 2011).

Whole-genome sequencing and sequence alignment

DNA was extracted from the 19 muscle samples using the Wizard Genomic DNA Purification kit (Promega). Each purified DNA sample was assessed by agarose gel electrophoresis. DNA concentration was measured with a Nanodrop ND-100 instrument (Thermo Fisher Scientific). Sequencing libraries were prepared using TruSeq SBS v3-HS Kit (Illumina) and the whole-genome sequenced using a 2×100 bp paired-end approach on an Illumina HiSeq 2000. Sequence alignments were carried out using the Burrows-Wheeler Alignment tool (BWA-v0.6.1-r104)[21] with the *aln* option with default parameters for mapping reads to the UMD3.1 bovine reference genome[22]. Potential PCR duplicates were removed using the MarkDuplicates tools from the Picard package version 1.4.0[23]. Only properly paired reads with a mapping quality of at least 30 ($-q = 30$) were retained. The resulting BAM files were then used for all subsequent analyses.

RNA sequencing and sequence alignment

RNA extraction and sequencing was performed as previously described[24–26]. Briefly, after transfer to ice-cold RNeasy RLT lysis buffer (Qiagen), LT tissue samples were homogenized using a Precellys tissue homogeniser (Bertin Technologie). Total RNA was isolated using RNeasy Midi columns (Qiagen) and then treated with RNAse-free DNase I (Qiagen) for 15 min at room temperature according to the manufacturer's protocols. The concentration of total RNA was measured with a Nanodrop ND-100 instrument (Thermo Scientific) and the quality was assessed with an RNA 6000 Nano Labchip kit using an Agilent 2100 Bioanalyzer (Agilent Technologies). All 19 samples had an RNA integrity number (RIN) value

greater than eight.

The mRNA-Seq libraries were prepared using the TruSeq RNA Sample Preparation Kit (Illumina) according to the manufacturer's instructions. Briefly, Poly-A containing mRNA molecules were purified from 4 μg total RNA of each sample using oligo (dT) magnetic beads and fragmented into $150-400$ bp pieces using divalent cations at 94 °C for 8 min. The cleaved mRNA fragments were converted to double-stranded cDNA using SuperScript II reverse transcriptase (Life Technologies) and primed by random primers. The resulting cDNA was purified using Agencourt AMPure® XP beads (Beckman Coulter). Then, cDNA was subjected to end-repair and phosphorylation and subsequent purification was performed using Agencourt AMPure® XP beads (Beckman Coulter). These repaired cDNA fragments were 3'-adenylated producing cDNA fragments with a single 'A' base overhung at their 3'-ends for subsequent adapter-ligation. Illumina adapters containing indexing tags were ligated to the ends of these 3'-adenylated cDNA fragments followed by two purification steps using Agencourt AMPure®XP beads (Beckman Coulter). Ten rounds of PCR amplification were performed to enrich the adapter-modified cDNA library using primers complementary to the ends of the adapters. The PCR products were purified using Agencourt AMPure®XP beads (Beckman Coulter) and size-selected (200 ± 25 bp) on a 2% agarose Invitrogen E-Gel (Thermo Scientific). Libraries were then checked on an Agilent Technologies 2100 Bioanalyzer using the Agilent High Sensitivity DNA Kit and quantified by quantitative PCR with the QPCR NGS Library Quantification kit (Agilent Technologies). After quantification, three different tagged cDNA libraries were pooled in equal ratios and a final qPCR check was performed post-pooling. Each library pool was used for 2 x 100 bp paired-end sequencing on one lane of the Illumina HiSeq2000 with a TruSeq SBS v3-HS Kit (Illumina). After sequencing, the samples were demultiplexed and the indexed adapter sequences were trimmed using the CASAVA v1.8.2 software (Illumina). The quality of the raw sequence reads was assessed using FastQC and Qualimap[27].

The *Bos taurus* reference genome sequence was downloaded from Ensembl (release 91, Bos_taurus-.UMD3.1.dna.toplevel.fa). To align the reads to the assembled reference genome the STAR RNA-Seq (version 2.4.2a) aligner was used[28]. Default values were used for mapping except for the intron alignment (*alignIntronMin*: 20 and *alignIntronMax*: 500,000). Reads for each sample were mapped separately to the reference genome sequence. Only paired reads were retained for alignment. The number of paired-reads uniquely aligning to transcribed regions of each transcript was calculated for all genes of the annotated transcriptome. The transcript paired-read count was calculated as the number of unique paired-reads that aligned within the exons of each transcript, based on the coordinates of mapped reads.

SNP identification and annotation

SNPs were called following the best practices from GATK (version 3.4-46) with HaplotypeCaller for DNA and RNA sequence data respectively[29,30]. First, reads were subjected to local realignment, coordinate sorting, base quality score recalibration and indel realignment. We then performed SNP and indel discovery and genotyping. In the GATK analysis, we used a minimum confidence score threshold of Q30 with default parameters. We also used multi-sample variant calling in order to distinguish between a homozygous reference genotype and a missing genotype among the analysed samples. SNPs were annotated with VEP[31] using the transcript set from ENSEMBL 87.

Detection of ASE SNPs

We used ASEReadCounter[9] to calculate read counts per allele. We performed an N-masking (replacing for each identified variant the nucleotide of the bovine genome reference sequence by N) to remove mapping bias and we only kept overlapping heterozygous SNPs from DNA and RNA to remove discordant genotypes, possibly due to imprinting or RNA editing. We only kept candidates with minimum 10 reads for at least one allele. To determine if the imbalance was significant, we used a binomial test against an allelic ratio of 0.5 with a p-value of 5% (Python).

Correlation analysis

The SNP being tested for ASE might not be the variant regulating the expression of the gene. So in order to determine the SNPs within the regulatory regions or potentially the regulatory variant itself, we detected SNPs in linkage disequilibrium with our ASE SNPs using PLINK 1.9[32] (intra-chromosomal analysis and $r^2 \geq 0.75$). We used HTSeq-count[33] to determine the number of reads for each transcript per individual and normalised this using DESeq2[34]. We computed the Spearman's rank correlation coefficient between the genotypes of ASE SNPs or SNPs in LD and expression level of the corresponding transcript. We performed a correction for multiple testing, for the same transcript, using the Bonferroni correction (Python).

ASE SNP validation

First-strand cDNA was synthesised from 500 ng of DNase I-treated total RNA using the SuperScript III First-Strand Synthesis System kit (Thermo Fisher Scientific) and oligo-dT primers with random hexamers, according to the manufacturer's instructions, in a total volume of 20μl. The resulting cDNA was diluted 1:10.

PCR and Pyrosequencing primers were designed using PyroMark Assay Design 2.0 (Qiagen) with sequences previously masked with RepeatMasker[35]. One of the forward or the reverse PCR primer had a 5'-biotin modification and was HPLC-purified. Primers were synthesized by IDT and are listed in Table S1. Polymerase chain reactions were performed in $50\mu l$, using $1\mu l$ of diluted cDNA or 100 ng of genomic DNA, 1 U GoTaq DNA polymerase (Promega), 1X PCR buffer, 1.5 mM MgCl2, $200\mu M$ of each dNTP and $0.3\mu M$ of each PCR primer. The following touchdown cycling protocol was used: 95 °C for 2 min, followed by 13 cycles of 95 °C for 1 min, 1 min of annealing (the annealing temperature was progressively lowered from 68 to 56 °C in steps of 1 °C every cycle) and 72 °C for 1 min 30 s. These initial cycles were followed by 20 cycles of 95 °C for 1 min, 55 °C for 1 min and 72 °C for 1 min 30 s, and a final extension step at 72 °C for 10 min. To check the quality of the amplification, 10 μl of PCR products were then analysed by gel electrophoresis with a 1% agarose gel.

Biotinylated PCR products ($20\mu l$) were immobilized on streptavidin-coated Sepharose beads (GE Healthcare), purified, washed and denatured using a 0.2 M NaOH solution and rewashed all using the PyroMark Vacuum workstation (Qiagen) as recommended by the manufacturer. Purified single-stranded PCR product was annealed to the pyrosequencing primer (diluted to $0.3\mu M$) and then sequenced using the PyroMark Q24 system (Qiagen), following the manufacturer's instructions. For validating candidate ASE SNPs, DNA and RNA (cDNA) from each sample were pyrosequenced simultaneously. The proportions of individual alleles for each SNP were obtained using the PyroMark Q24 software version 1.0.10 (Qiagen). Genomic DNA was examined to confirm the heterozygosity. The final ASE ratio for each SNP of each sample analysed was calculated using the following formula: ASE ratio = (allele 1 %/ allele 2 %) RNA / (allele 1 %/ allele 2 %) genomic DNA.

Prediction of microRNA binding sites

Prediction of micro RNA (miRNA) binding sites was done as follows: first, for SNPs within 3'UTR regions, flanking sequences ($\pm$100 bases) were retrieved using the whole-genome reference sequence (UMD3.1). Then, we created two versions of this sequence, one with the reference allele and one with the alternate allele. Next, we used miRanda[36] for both sequences with all known bovine miRNAs, using the default parameters. Bovine miRNA sequences were retrieved from the miRBase database (version 21). To finish, we selected miRNAs which could bind only one of these two sequences.

Data Availability

Whole-genome sequences were submitted to SRA (submission in progress). RNA-Seq data analysed during the current study is available from the European Nucleotide Archive (accession numbers ERP002220, E-MTAB-2646, E-MTAB-4625 and E-MTAB-6947). The ASE SNPs identified in this study are included in the Table S4.

Results and Discussion

DNA and RNA sequencing data statistics

Sequencing of all 19 whole-genome sequences generated a total of 5.3 billion of raw paired-end reads corresponding to 537.51 Gb. Approximately, 92 to 400 million pairedend reads were obtained for each library. On average, 83% (56-92%) of the paired-end reads were properly aligned with BWA on the UMD3.1 bovine reference genome (Table S2).

Sequencing of all 19 RNA-Seq libraries generated a total of 1.4 billion raw paired-end reads. Approximately, 35 to 180 million paired-end reads were obtained for each library. On average, 89% (86–91%) of the reads were uniquely mapped (Table S3). In a previous study[26], 17 of our 19 RNA samples were sequenced and mapping was performed using BWA (version 0.5.9-r16)[21]. 63–76% of the mapped reads were aligned. The increase of the mapping rate (on average 17.8% more reads) indicates that STAR performs best. This is largely because STAR is a splice aware aligner. The mapping performance is comparable to other studies done in cattle with STAR and the same reference genome (UMD3.1). For instance, 90% of transcripts from Holstein-Friesian peripheral blood leukocytes were mapped[37].

The count of transcripts was performed using HTSeq-count[33] and was normalized with DESeq2[34]. In our samples, we found 18,206 transcripts (corresponding to 16,338 genes) with an expression in at least 3 individuals among the 19.

Variant detection

We identified 11,943,766 and 269,390 single nucleotide variants (SNVs) from WGS and RNA-Seq data, respectively. We identified on average $11,344,542 \pm 7.12\%$ SNVs per individual from WGS and on average $53,732 \pm 31.85\%$ SNVs per individual from RNA-Seq reads. On average, 26.2% and 34.2% of the detected SNVs were heterozygous in WGS and RNA-Seq, respectively.

Among the SNVs identified from WGS (Table 1), we identified 8,099,157 (67.81%), 2,922,660 (24.47%), 413,619 (3.46%), 405,237 (3.39%) as intergenic, intronic, upstream gene, downstream gene variants, re-

spectively. We identified 69,096 (0.58%) exonic variants (56.62% synonymous, 43.32% missense and 0.07% coding sequence variants). For the other types of variants, the percentage was less than 0.20%: 19,332 3'UTR (0.16%) and 3,544 5'UTR variants (0.03%).

Among variants found with RNA-Seq data, we identified 54,410 (20.20%), 106,700 (39.61%), 14,734 (5.47%), 53,630 (19.91%) as intergenic, intronic, upstream gene, downstream gene variants, respectively. We identified 24,160 (8.97%) exonic variants (59.25% synonymous, 40.5% missense and 0.24% coding sequence variants).

We found 67.8% of SNPs from WGS data as intergenic. This percentage is in agreement with the 70.4% of the intergenic part of the bovine genome. This proportion is also similar in others studies done in cattle. For instance 73% of intergenic, 26.2% of intronic, 4.26% of downstream gene and 4.14% of upstream gene variants were found in Hanwoo and Yanbian cattle[38] or 65.6% of intergenic and 33.6% were identified of intronic variants in Qinchuan cattle[39]. Interestingly, we found 20.20% (54,410) of SNPs identified from our RNA-Seq data as intergenic. These SNPs could be located in transcripts of large intergenic non-coding RNAs. Indeed, we found 7,706 (14,16%) intergenic SNPs from our RNA-Seq data within lincRNAs previously identified from six of our samples by Billerey and collaborators[25]. We also found 39.61% of SNPs identified from our RNA-Seq data in intronic regions. These SNPs could be from premature transcripts (before splicing).

RNA-Seq and DNA-Seq SNP comparison

We compared SNPs detected from WGS with SNPs from RNA-Seq data for each individual. On average, we detected 11,306,326 SNPs only from WGS (out of 11,943,766 detected SNPs), 15,516 SNPs only from RNA-Seq reads (out of 269,390 detected SNPs), and 38,217 of the SNPs from both (Table 2). We focused on overlapping SNPs identified from WGS and RNA-Seq data and checked the concordance between their genotype. This overlap is on average 90% (75.7% to 96.0%) concordant (69% for both homozygous and 31%for both heterozygous). For the 10% discordant SNPs, 84.3% are homozygous from DNA-Seq and heterozygous from RNA-Seq data. This could be explained by RNA editing. 15.7% are heterozygous from DNA-Seq and homozygous from RNA-Seq; this could be explained by gene imprinting (mono-allelic expression). Alternatively, discrepancies between DNA and RNA genotypes could be due to sequencing errors. To study the allelic imbalance, we only kept the heterozygous concordant SNPs.

	DNA		RNA	
Variant consequences	Number of snps	%	Number of snps	%
intergenic_variant	8,099,157	67.81%	54,410	20.20%
intron_variant	2,922,660	24.47%	106,700	39.61%
upstream_gene_variant	413,619	3.46%	14,734	5.47%
downstream_gene_variant	405,237	3.39%	53,630	19.91%
synonymous_variant	39,119	0.33%	14,315	5.31%
missense_variant	29,931	0.25%	9,786	3.63%
3_prime_UTR_variant	19,332	0.16%	11,555	4.29%
splice_region_variant	6,471	0.05%	475	0.18%
non_coding_exon_variant	3,930	0.03%	0	0.00%
5_prime_UTR_variant	3,544	0.03%	1,374	0.51%
Unindentified	269	0.00%	132	0.05%
splice_donor_variant	153	0.00%	73	0.03%
splice_acceptor_variant	148	0.00%	44	0.02%
initiator_codon_variant	62	0.00%	0	0.00%
coding_sequence_variant	46	0.00%	59	0.02%
mature_miRNA_variant	37	0.00%	0	0.00%
stop_retained_variant	32	0.00%	15	0.01%
non_coding_transcript_variant	19	0.00%	11	0.00%
frameshift_variant	0	0.00%	1,221	0.45%
protein_altering_variant	0	0.00%	1	0.00%
non_coding_transcript_exon_variant	0	0.00%	855	0.32%

Table 1. Summary of SNPs detected in RNA and DNA with their annotation frequencies.

Individual	DNA only	RNA only	Overlap	BH	Bh	Concordant	Hh	hH	Discordant
LIM1	11,420,182	19,039	44,861	27,410	11,354	86.4%	4,979	1,118	13.6%
LIM2	11,549,679	17,681	46,624	29,535	12,671	90.5%	3,974	444	9.5%
LIM3	11,753,420	15,867	49,721	31,024	16,413	95.4%	1,633	651	4.6%
LIM4	11,770,633	13,801	38,579	23,198	12,968	93.7%	1,149	1,264	6.3%
LIM5	11,668,108	11,596	36,346	22,687	11,637	94.4%	1,513	509	5.6%
LIM6	11,645,235	16,568	44,888	27,925	12,860	90.9%	3,295	808	9.1%
LIM7	11,287,139	6,218	15,075	9,088	3,439	83.1%	1,947	601	16.9%
LIM8	11,734,961	18,876	55,713	35,061	17,430	94.2%	2,306	916	5.8%
LIM9	11,563,319	13,215	33,473	21,119	9,012	90.0%	2,897	445	10.0%
LIM13	8,718,858	27,165	28,651	18,020	3,671	75.7%	6,707	253	24.3%
LIM14	11,665,886	12,410	34,686	22,388	9,932	93.2%	1,796	570	6.8%
LIM15	11,516,569	15,344	40,398	25,775	10,135	88.9%	3,931	557	11.1%
LIM16	11,766,765	12,041	35,918	22,612	11,854	96.0%	890	562	4.0%
LIM17	9,511,239	21,194	28,415	17,675	3,677	75.1%	6,863	200	24.9%
LIM18	11,755,926	8,686	24,893	15,029	8,585	94.9%	902	377	5.1%
LIM19	11,517,295	15,901	40,528	25,083	11,315	89.8%	3,573	557	10.2%
LIM20	11,330,071	12,058	19,755	12,190	4,423	84.1%	2,753	389	15.9%
LIM21	11,110,581	14,100	30,031	19,059	6,466	85.0%	4,147	359	15.0%
LIM22	11,534,319	23,041	77,560	45,999	24,815	91.3%	5,907	839	8.7%
Average	11,306,326	15,516	38,217	23,730	10,666	89.1%	3,219	601	10.9%

Table 2. Distribution of detected SNPs from RNA-Seq and WGS data per individual. BH: Both Homozygous. Bh: Both Heterozygous. Concordant: Rate of BH and Bh. Hh: Homozygous in DNA and Heterozygous in RNA. hH: Heterozygous in DNA and Homozygous in RNA. Discordant: Rate of Hh and hH.

ASE SNP identification

Using ASEReadCounter we calculated reads count per allele for all heterozygous concordant SNPs from alignment to the UMD3.1 reference genome sequence and the N-masked genome sequence. On average, the N-masking removed 27.1% of the candidate SNPs from ASE detection. We identified 6,908 ASE SNPs (Table S4) in 2,451 genes corresponding to 9.8% of all bovine genes (25,066), 15% of the genes with detectable expression in *Longissimus thoraci* muscle (16,338) and 20% of the genes with at least one heterozygous SNPs (12,269). On average, we detected 574 ASE SNPs per individual (min: 184, max: 991) corresponding to 3.2% of the heterozygous SNPs from RNA-Seq data (Table S5). Last, we removed ASE SNPs within CNV regions previously identified within our Limousine animals[40] and kept 5,658 ASE SNPs located in 2,119 genes.

We then checked the distribution of the ASE SNPs across chromosomes. There is a weak correlation between the number of ASE SNPs per chromosome and the size of the chromosomes ($\rho = 0.45$, p-value$= 0.015$). However, the number of ASE SNPs per chromosome is strongly correlated with the number of coding genes ($\rho = 0.84$, p-value$= 9.13E-09$) and with the number of expressed genes ($\rho = 0.85$, p-value$= 4.81E-09$) (Figure 1).

We compared our detected ASE SNPs with ASE SNPs previously identified by Chamberlain and collaborators in a Holstein muscle sample[19]. In their study, ASE detection was performed on one lactating dairy cow using TOPHAT2[41] for the read alignment and a Chi-squared test. We found 118 ASE SNPs in common with the 2,006 ASE SNPs from Holstein muscle representing 5.9% of their detected ASE SNPs. We investigated why we do not detect the remaining ASE SNPs in our results. 684 of these SNPs (34.1%) were not polymorphic in our Limousine animals, 43 others SNPs (2.1%) are not showing heterozygosity among our 19 individuals and 38 SNPs (1.9%) are located on the chromosome X (excluded because we have only males). For the 1,123 remaining ASE SNPs (60.0%) identified in Holstein muscle, we found at least one heterozygous Limousine animal. This discrepancy might be due to differences in ASE detection methods or in breed gene regulation.

Functional annotation of ASE SNPs and of their genes

4,193 of the detected ASE SNPs were located within cattle QTL regions reported in Animal QTLdb[42] (Table S6). Interestingly, 1,213 of these ASE SNPs were inside QTL regions found in Limousine and 2,107 of these SNPs were in QTL regions linked to growth or meat traits.

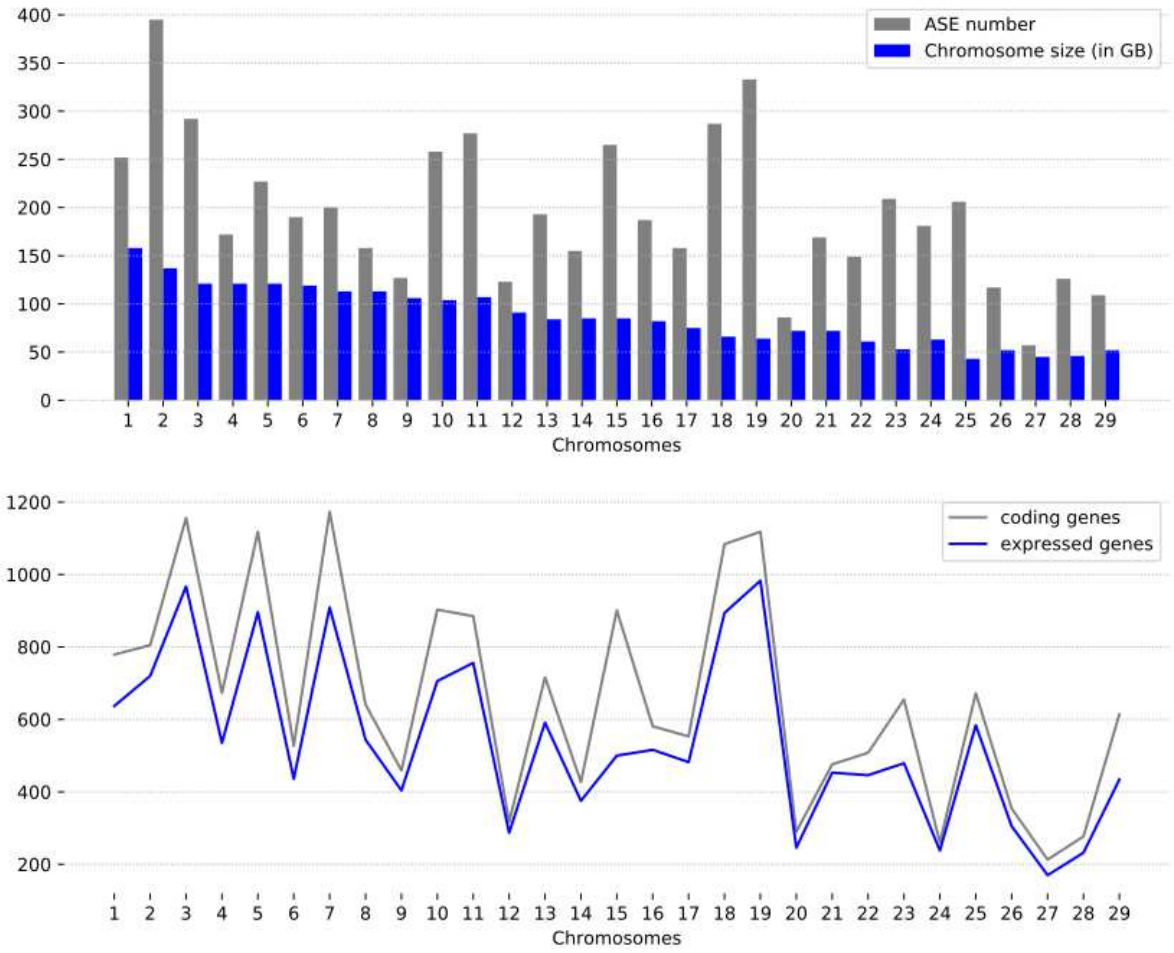

Figure 1. Chromosomal distribution with the number of ASE SNPs (grey bars), the size of the genomes (blue bars), the number of genes: total (blue line) and only expressed in muscle (grey line).

In order to study the impact of genes affected by ASE on specific biological pathways, we performed a Gene Ontology (GO) enrichment. This analysis was carried out by first converting the cow gene list into a human gene list using Biomart[43]. This resulted in a list of 2,143 genes that was tested for enriched GO terms using the GOrilla tool[44] with a background gene list of all expressed genes in *Longissimus thoraci* muscle (13,998).

In total, the genes showing ASE corresponded to 127 enriched functions (q-value < 0.05), with many of these related to striated muscle development (Table S7). The top 20 most-enriched terms are presented in figure 2. Thirteen functions were related to muscle functions or components: contractile fiber part (GO:0044449), Z disc (GO:0030018), actin binding (GO:0003779), actin filament-based process (GO:0030029), cytoskeletal protein binding (GO:0008092), muscle contraction (GO:0006936), muscle system process (GO:0003012), structural constituent of muscle (GO:0008307), actin filament binding (GO:0051015), muscle alpha-actinin binding (GO:0051371), sarcomere organisation (GO:0045214) and M band (GO:0031430). The seven GO terms not directly related to muscle were linked to intracellular part and/or organelle and can be associated with contractile fibre part, mitochondrion or nucleus.

ASE validation

We used Pyrosequencing in order to validate ASE SNPs. Several filters were applied to narrow down the number of ASE SNPs to test. Firstly, we kept ASE SNPs present in a QTL region associated with growth or meat quality traits reported in Animal QTLdb[42]. Secondly, we removed SNPs absent from dbSNP. Then, we only kept ASE SNPs present in exonic, 5'UTR or 3'UTR regions. Finally, we selected two ASE SNPs located within *CAST* and we choose randomly four extra ASE SNPs.

We tested these 6 ASE SNPs by Pyrosequencing with replicates (Table 3). Technical replicates obtained from independent experiments show standard deviations ranging from $0-4\%$, indicating that our Pyrosequencing procedure has negligible inter-PCR and Pyrosequencing variations. The allele frequencies determined for genomic DNA samples, which we analysed in duplicate showed an average variation of $2\% \pm 1\%$ (n=4). For the cDNA samples, the average variation between replicates was $2\% \pm 2\%$ (n=4). We could therefore detect allele frequency differences larger than 4%. Five ASE SNPs were validated by Pyrosequencing. For example, we observed, for the validated ASE SNPs rs110694123 in *PALLD* gene, 47% for allele G (complementary base of C) and 53% for allele A (complementary base of T) in gDNA and we observed 33% and 67% in cDNA (Figure 3). We get an ASE ratio of 1.80 showing an allelic

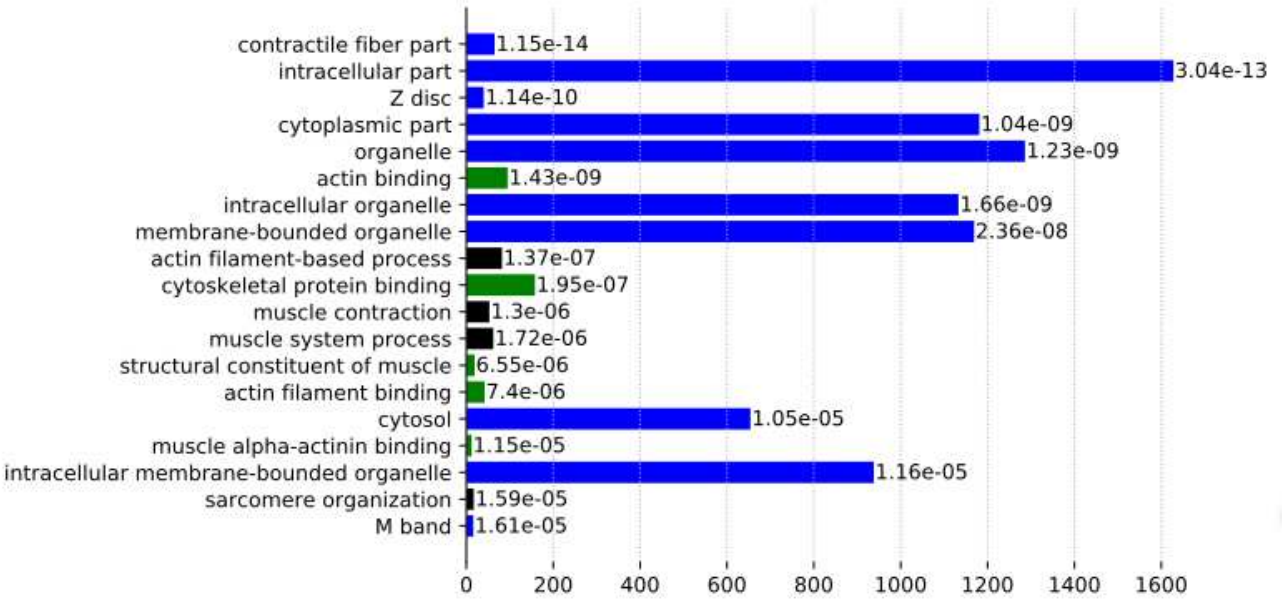

Figure 2. Enriched GO terms for genes affected by ASE. Functional enrichments for gene ontology (GO) terms associated with the 2,119 genes affected by ASE SNPs (5,658). Only the top ranked 20 terms are shown. The horizontal bar represents the number of ASE-genes involved, with the corresponding q-values. The GO terms categories included Biological Process (black), Cell Component (blue) and Molecular Function (green). The enrichment analysis was performed with the GOrilla tool.

Validated	BTA	Position	ID	REF	ALT	ASE count	Gene	Annotation
Yes	3	32,003,949	rs382378456	C	A	407/336	*ATP5F1*	3'UTR variant
No	7	5,520,428	rs208775256	G	C	26/12	*PGLS*	missense variant
Yes	7	98,579,574	rs41255587	G	A	146/208	*CAST*	3'UTR variant
Yes	7	98,580,401	rs209641420	A	C	303/221	*CAST*	3'UTR variant
Yes	8	572,167	rs110694123	G	A	48/73	*PALLD*	synonymous variant
Yes	8	944,049	rs109919583	C	T	47/121	*CBR4*	3'UTR variant

Table 3. ASE SNPS tested by Pyrosequencing. REF: reference allele, ALT: alternative allele, ASE count: number of reference allele reads/number of alternative allele reads.

imbalance in favour of allele A (it means there is 1.80 more expression of transcripts with the A allele than with the G allele). This is consistent with the ASE ratio computed from the read counts for this SNP (1.52 with 39.67% for G and 60.33% for A).

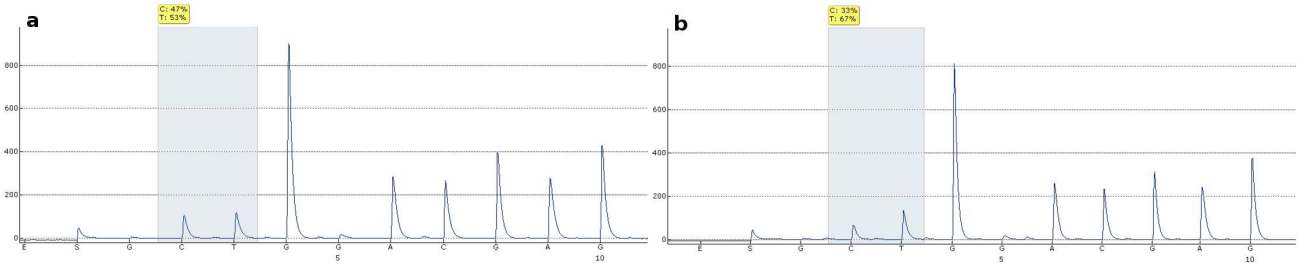

Figure 3. Pyrosequencing results of one ASE-SNP in *PALLD* gene. (a) In gDNA, 47% for allele C and 53% for allele T. (b) In cDNA, 33% for allele C and 67% for allele T.

Cis-regulation of genes showing allele specific expression

Our detected ASE SNPs are probably not the causative variants, but rather markers in *cis* with the causative polymorphisms. It is known that the majority of causative SNPs are in regulatory regions instead of coding regions[45]. Therefore, we were looking for a link between ASE SNPs and the putative causative SNPs in *cis*. With this in mind, we used PLINK to identify all the SNPs in linkage disequilibrium (LD) ($r^2 \geq 0.75$) with our predicted ASE SNPs. We obtained 2,955 SNPs (including ASE SNPs) with genotypes for all the 19 individuals. For each transcripts showing allele specific expression, we calculated the Spearman correlation coefficient score between expression level of genes containing ASE SNPs and genotypes of SNPs in LD with ASE SNPs. We computed correlations between 2,794 SNP genotypes and 1,085 unique transcripts, averaging 2.74 SNP genotypes per transcript (min 1, max 37). We found 100 significant correlations with 45 transcripts ($\rho > |0.6|$ and q-value < 0.05) including 42 negative correlations (Table S8). 25 of those correlations involved an ASE SNP.

For example, we found one SNP (C/T, rs41691181) in LD ($r^2 = 0.79$, distance of 12.5 kb) with a SNP (C/T, rs208256739) in upstream and exonic (synonymous variant) regions of *APMAP* respectively. The second SNP shows ASE in one individual (LIM8) among the nineteen. The genotypes of the first SNP (8 C/C, 7 C/T, 4 T/T) is significantly correlated ($\rho - 0.75$ and q-value$= 0.000188$) to the *APMAP* level expression. Indeed, we found on average for the 19 animals 404, 323 and 214 transcripts (read counts) for C/C, C/T and T/T animals (Figure 4a) showing an expression bias in favour of the C allele.

We investigated how this SNP (rs41691181) in the upstream gene region could cause this allelic imbal-

ance by testing if the SNP could alter Transcription Factor Binding site (TFBS) using TFBS-match[46] with the SNP flanking sequences (±10 bases). None of the allele-specific sequences of these SNPs were located in predicted TFBS.

We extended the TFBS search for 5 other SNPs in 5 different genes (*5S_rRNA*, *LRRC66*, *ENSBTAG00000026637*, *GLOD4* and *PLK1*) with a significant correlation in the upstream region without detecting any TFBS.

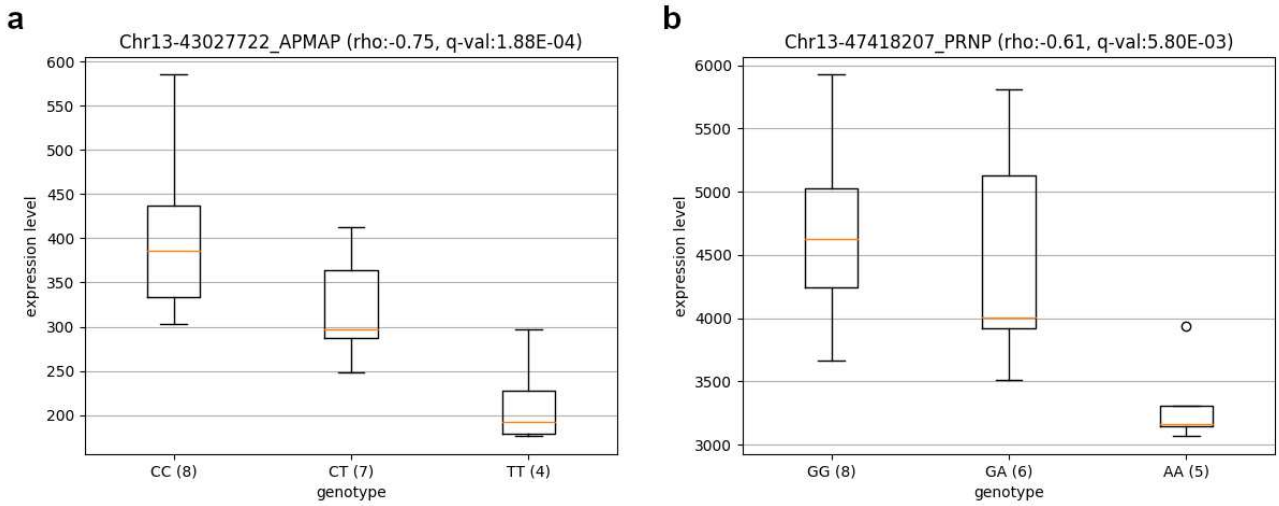

Figure 4. Boxplots of SNP showing genetic variations of *APMAP* **(a)** or *PRNP* **(b)** expressions. (N) number of animals per genotype.

In another example, we found one SNP (G/A, rs109763272) in LD ($r^2 = 0.86$, distance of 274 bases) with a SNP (G/A, rs378125518). Both SNPs are in 3'UTR region of the *PRNP* gene and show ASE in four individuals among the nineteen. The genotypes of the first SNP (8 G/G, 6 G/A, 5 A/A) is significantly correlated ($\rho - 0.61$ and q-value$= 0.0057966$) to the *PRNP* expression level. On average, the PRNP expression level was 4,641 transcripts for G/G individuals, 4,455 for G/A individuals and 3,324 for A/A individuals (Figure 4b) showing an expression bias in favour of allele G. Given that this correlated SNP is also an ASE SNP, we looked if allele counts estimated with ASEReadCounter is in agreement with the transcript expression level. Indeed, transcripts with the G allele are 1.54 times more expressed than transcripts with the A allele.

We investigated how this SNP (rs109763272) in 3'UTR region could cause this allelic imbalance. It is known that polymorphisms in microRNA (miRNA) binding sites may affect miRNA/target gene interaction[47]. Therefore, we used miRanda to detect miRNA binding sites within this SNP flanking region.

We predicted 9 miRNAs which could bind the reference allele (G) and 5 miRNAs which could bind the alternate allele (A) (Table S9). Interestingly, we noticed less expression with the alternate allele (Figure 4b). This could suggest that some of the 5 detected miRNAs binding with the A allele could reduce the expression of *PRNP*.

We lack data on miRNAs expression in our samples, but several studies describing catalogs of miRNAs expressed in bovine muscle or skeletal muscle satellite cells have been published[48–58]. However, no study describes so far miRNAs expressed in Limousin animals. We found that all fourteen miRNAs impacted by the SNP rs109763272 are expressed in muscle[50–53], including in *Longissimus dorsi/thoracis*[53] (Table S10). We therefore cannot exclude any of the 5 miRNAs binding to the A allele, or any of the 9 miRNAs binding to the G allele as candidate *PRNP* regulators. Further work is needed to identify which of these candidate miRNAs reduce *PRNP* expression level.

We extended the miRNA binding sites prediction analysis to all SNPs with a significant correlation and located in a 3'UTR region (Table S9). We analysed 13 additional SNPs present in 6 other genes (1 SNP in *ANKRD*, 1 in *CCDC90B*, 2 in *FAM32A*, 2 in *TYK2*, 3 in *IMP3* and 4 in *TTC3*). We found no binding sites for 3 of these SNPs and for the remaining 10 SNPs we always found allele specific binding sites for both alleles (Figure S1) including 8 SNPs with a lower expression with the alternate allele. This could suggest that some of the detected miRNAs are binding with the alternate allele to reduce the gene expression. We found 2 SNPs with a lower expression of the reference allele. Similar to the alternate allele, the detected miRNAs binding with the reference allele could reduce gene expression. Survey of miRNAs expressed in bovine muscle, allowed us to excluded only eleven miRNAs (Table S10). Further work is needed to identify which SNPs impact target sites of the remaining 386 miRNAs.

For most of the 45 genes for which we had a significant correlation between expression level and SNP (ASE SNP or SNP in LD with an ASE SNP) genotypes we couldn't find SNPs altering TFBSs or the binding sites of miRNAs. It is therefore likely that epigenetic mechanisms might also play a role, rather than just *cis*-regulatory genetic variants (in TFBS or 3'UTR).

ASE genes potentially involved in meat quality traits

The aldehyde oxydase 1 (*AOX1*) gene encodes a homodimeric protein, which produces hydrogen peroxide. In mouse, it is involved in myogenesis[59]. Therefore, it might play a role in muscle development in cattle. We detected eleven ASE SNPs in this gene with six also detected by Chamberlain and collaborators[19]. Among these 6 ASE SNPs, three had genotypes significantly correlated to the expression of this

gene. In addition, we found 13 others SNPs in *AOX1* with significant correlation (Figure S2).

The palladin (*PALLD*) gene encodes a cytoskeletal associated protein, which exists as multiple isoforms[60]. This actin associated protein plays a significant role in regulating cell adhesion and cell motility. It is also important for the early smooth muscle cell differentiation in mouse[61]. In cattle, palladin might play dual roles (positive and negative) in maintaining the proper skeletal myogenic differenciation[62]. We detected two ASE SNPs in this gene including one experimentally validated by Pyrosequencing. Interestingly, these SNPs are within a QTL region associated with average daily gain (ADG) trait in Hereford[63].

The calpastatin (*CAST*) gene encodes an inhibitor of protease μ-calpain, which has a known effect on beef muscle tenderness variation[64]. Interestingly, a more recent study confirmed that *CAST* affected meet tenderness in *Longissimus* muscle in Limousine crossed-breed animals[65]. We detected seven ASE SNPs in this gene including two experimentally validated.

These 3 genes could be associated with meat quality and carcass traits. Interestingly, one of the ASE SNPs found in *AOX1* is a missense variant. This SNP (rs109201304) modifies a glycine residue into a cysteine amino acid and is located within a protein region conserved in mammals (Fig. S3). This residue (p.G1023C) lies within the substrate pocket subdomain IV of the large C-terminal domain which is important for substrate access and positioning but also in the dimerization of the two *AOX1* monomeric subunits66,67. Several studies performed on *AOX1* variants resulting from rat or human missense SNPs have shown that some of these SNPs increased or decreased the rate of superoxide radical production. Further work is needed to investigate whether r109201304 can affect the catalytic activity of bovine AOX1. We didn't find any missense polymorphisms in *PALLD* and *CAST* but we identified several synonymous variants (2 in *PALLD* and 2 in *CAST*). They don't alter the primary sequence of the corresponding proteins however it has been shown that codon usage can vary between genes and that this codon bias can affect RNA secondary structure, splicing and translation. Further work is needed to investigate the phenotypic impact of these variants/genes.

Biological relevance of allele specific expression in muscle

Overall we identified 5,658 ASE SNPs in 13% of genes (2,119) with detectable expression in *Longissimus thoracis* muscle. The high number of genes potentially impacted by allele-specific imbalance, prompted us to investigate if some of these ASE SNPs could have a major impact on muscle biology.

First we looked if ASE SNPs could induce a gene loss-of-function. We didn't find any ASE SNP that could create or remove stop codons and causing consequently protein truncations or changes in the open

reading frame, respectively. However, we identified 14 ASE SNPs that according to the VEP annotation have or could perturb the splicing of the corresponding gene. Further work is needed to check this potential impact.

Second we investigated further the 421 missense ASE SNPs. According to the VEP annotation, only 37 of those missense ASE SNPs are predicted to be deleterious. 95% of these deleterious ASE SNPs are found in only one or two animals. Interestingly, we found one T/C deleterious ASE SNP (chromosome 10, position 37,912,737) within *ZFP106*, in one animal (LIM18). *ZFP106* encodes a zinc fingered RNA-binding protein. Disruption of *Zfp106* in mice induces several skeletal muscle phenotypic abnormalities[66–68], such as severe muscle wasting[67], loss of muscle strength[66–68] and degeneration of muscle fibers[68] in homozygous knock out *Zfp106* -/- mice. Heterozygous *Zfp106* +/- mice are comparable to wild type littermates[67,68]. These results suggest that *ZFP106* might not be a dosage-sensitive gene and that haploinsufficiency of ZFP106 (in ASE SNP heterozygous animals) might not impact muscle physiology.

We also found a deleterious ASE SNP (rs110365838) within *MAP4*, a muscle-specific microtubule associated protein which is expressed in early myogenesis[69] and that is required for muscle cell differentiation[70]. This ASE SNP was detected in two animals (LIM2 and LIM15). We didn't find, so far, any information on potential consequences of deleterious variants within this gene. However, because of the critical role of MAP4 in muscle development, it will be interesting to investigate if the two heterozygous animals for this ASE SNP have normal amount of MAP4 protein.

Third, we examined if ASE SNPs could impact genes important for muscle cell development or function. We focused on ASE SNPs located in downstream, upstream, 5' or 3' UTR regions, as they might have an effect on the regulation of the transcription of important genes. We found that *myogenin* (MYOG), a muscle-specific transcription factor required to induce myogenesis[71], had in total 21 ASE SNPs, including 5 and 7 in downstream and 3'UTR regions, respectively. However, disruption of murine *myogenin* showed no overt effects in heterozygous *Myog* +/- mice[72] suggesting that a potential reduction of MYOG in animals heterozygous for those 12 ASE SNPs might not have phenotypic consequences.

Conclusion

We performed a genome-wide survey of ASE using 19 Limousine muscle samples combining WGS and RNA-Seq data. This analysis shows that ASE is pervasive in beef muscle. We identified 5,658 ASE SNPs located in 2,119 genes and 37.2% of these ASE SNPs are found within QTLs associated to meat or

carcass traits. We validated 5 out of 6 selected ASE SNPs suggesting that our pipeline identify mostly true ASE SNPs. In addition, we detected SNPs with genotypes significantly associated with gene expression levels. For example, we identified one SNP in the 3'UTR region of *PRNP* that could be a causal mutation by modifying binding sites of several miRNAs. We showed that our *in silico* ASE approach can facilitate the identification of candidate *cis*-regulatory SNPs. However, further work is needed to validate these candidates. In the future, functional analyses of the impact of polymorphisms within TF or miRNA binding sites will try to elucidate the molecular mechanisms underlying gene expression imbalance.

References

1. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
2. Amit, I. *et al.* Unbiased reconstruction of a mammalian transcriptional network mediating the differential response to pathogens. *Sci.* **326**, 257–263 (2009).
3. Haley, C. & De Koning, D. J. Genetical genomics in livestock: potentials and pitfalls. *Animal Genet.* **37**, 10–12 (2006).
4. Zou, F. *et al.* Brain expression genome-wide association study (egwas) identifies human disease-associated variants. *PLOS Genet.* **8(6)**, e1002707 (2012).
5. Sabbagh, U., Mullegama, S. & Wyckoff, G. J. Identification and evolutionary analysis of potential candidate genes in a human eating disorder. *BioMed Res. Int.* **2016**, 1–11 (2016).
6. Grigoryev, D. N. *et al.* Identification of new biomarkers for acute respiratory distress syndrome by expression-based genome-wide association study. *BMC Pulm. Medicine* **15**, 95 (2015).
7. The GTEx Consortium. The genotype-tissue expression (gtex) project. *Nat. Genet.* **45**, 580–585 (2013).
8. Lopdell, T. J. *et al.* Dna and rna-sequence based gwas highlights membrane-transport genes as key modulators of milk lactose content. *BMC Genomics* **18**, 968 (2017).
9. Castel, S. E. *et al.* Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
10. Muráni, E., Ponsuksili, S., Srikanchai, T., Maak, S. & Wimmers, K. Expression of the porcine adrenergic receptor beta 2 gene in longissimus dorsi muscle is affected by cis-regulatory dna variation. *Animal Genet.* **40**, 80–89 (2009).
11. Chen, J. *et al.* A uniform survey of allele-specific binding and expression over 1000-genomes-project individuals. *Nat. Commun.* **7**, 11101 (2016).
12. Lagarrigue, S. *et al.* Analysis of allele-specific expression in mouse liver by rna-seq: A comparison with cis-eqtl identified using genetic linkage. *Genet.* **195**, 1157–1166 (2013).
13. Fear, J. M. *et al.* Buffering of genetic regulatory networks in drosophila melanogaster. *Genet.* **203**, 1177–1190 (2016).
14. Maroilley, T. *et al.* Deciphering the genetic regulation of peripheral blood transcriptome in pigs through expression genome-wide association study and allele-specific expression analysis. *BMC Genomics* **18** (2017).
15. Zhuo, Z., Lamont, S. J. & Abasht, B. Rna-seq analyses identify frequent allele specific expression and no evidence of genomic imprinting in specific embryonic tissues of chicken. *Sci. Reports* **7** (2017).

16. Ghazanfar, S. *et al.* Gene expression allelic imbalance in ovine brown adipose tissue impacts energy homeostasis. *PlOS ONE* **12**, e0180378 (2017).

17. Esteve-Codina, A. *et al.* Exploring the gonad transcriptome of two extreme male pigs with rna-seq. *BMC Genomics* **12**, 552 (2011).

18. Chitwood, J. L., Rincon, G., Kaiser, G. G., Medrano, J. F. & Ross, P. J. Rna-seq analysis of single bovine blastocysts. *BMC Genomics* **14**, 350 (2013).

19. Chamberlain, A. J. *et al.* Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics* **16**, 993 (2015).

20. Allais, S. *et al.* The two mutations, q204x and nt821, of the myostatin gene affect carcass and meat quality in young heterozygous bulls of french beef breeds. *J. animal science* **88**, 446–54 (2009).

21. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinforma.* **25**, 1754–1760 (2009).

22. Zimin, A. V. *et al.* A whole-genome assembly of the domestic cow, bos taurus. *Genome Biol.* **10**, R42 (2009).

23. Picard tools by broad institute. http://broadinstitute.github.io/picard/.

24. Djari, A. *et al.* Gene-based single nucleotide polymorphism discovery in bovine muscle using next-generation transcriptomic sequencing. *BMC Genomics* **14**, 307 (2013).

25. Billerey, C. *et al.* Identification of large intergenic non-coding rnas in bovine muscle using next-generation transcriptomic sequencing. *BMC Genomics* **15**, 499 (2014).

26. Meersseman, C. *et al.* Genetic variability of the activity of bidirectional promoters: a pilot study in bovine muscle. *DNA Res.* **24**, 221–33 (2017).

27. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinforma.* **32**, 292–294 (2016).

28. Dobin, A. *et al.* Star: Ultrafast universal rna-seq aligner. *Bioinforma.* **29**, 15–21 (2013).

29. McKenna, A. *et al.* The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

30. Auwera, G. A. *et al.* From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1–11.10.33 (2013).

31. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

32. Chang, C. C. *et al.* Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 1–16 (2015).

33. Anders, S., Pyl, P. T. & Huber, W. Htseq—a python framework to work with high-throughput sequencing data. *Bioinforma.* **31**, 166–69 (2015).

34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.* **15**, 550 (2014).

35. Smit, A., Hubley, R. & Green, P. Repeatmasker open-4.0. http://www.repeatmasker.org (2013-2015).

36. Enright, A. J. *et al.* Microrna targets in drosophila. *Genome Biol.* **5**, R1 (2004).

37. McLoughlin, K. E. *et al.* Rna-seq transcriptional profiling of peripheral blood leukocytes from cattle infected with mycobacterium bovis. *Front. Immunol.* **5**, 396 (2014).

38. Choi, J.-W. *et al.* Whole-genome resequencing analysis of hanwoo and yanbian cattle to identify genome-wide snps and signatures of selection. *Mol. Cells* **38**, 466–473 (2015).

39. Xu, Y. *et al.* Whole-genome sequencing reveals mutational landscape underlying phenotypic differences between two widespread chinese cattle breeds. *PLOS ONE* **12**, e0183921 (2017).

40. Letaief, R. *et al.* Identification of copy number variation in french dairy and beef breeds using next-generation sequencing. *Genet. Sel. Evol.* **49**, 77 (2017).

41. Kim, D. *et al.* Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).

42. Hu, Z.-L., Park, C. A. & Reecy, J. M. Developmental progress and current status of the animal qtldb. *Nucleic Acids Res.* **44**, D827–D833 (2015).

43. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).

44. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinforma.* **10**, 48 (2009).

45. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory dna. *Sci.* **337**, 1190–1195 (2012).

46. Kel, A. *et al.* Matchtm: a tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res.* **31**, 3576–3579 (2003).

47. Vymetalkova, V. *et al.* Polymorphisms in microrna binding sites of mucin genes as predictors of clinical outcome in colorectal cancer patients. *Carcinog.* **38**, 28–39 (2017).

48. Muroya, S. *et al.* Profiling of differentially expressed microrna and the bioinformatic target gene analyses in bovine fast- and slow-type muscles by massively parallel sequencing1. *J. Animal Sci.* **91**, 90–103 (2013).

49. Miretti, S., Volpe, M. G., Martignani, E., Accornero, P. & Baratta, M. Temporal correlation between differentiation factor expression and micrornas in holstein bovine skeletal muscle. *Animal* **11**, 227–235 (2017).

50. Zhang, W. W. *et al.* Effect of differentiation on microrna expression in bovine skeletal muscle satellite cells by deep sequencing. *Cell. & Mol. Biol. Lett.* **21**, 8 (2016).

51. Sadkowski, T., Ciecierska, A., Oprzadek, J. & Balcerek, E. Breed-dependent microrna expression in the primary culture of skeletal muscle cells subjected to myogenic differentiation. *BMC Genomics* **19**, 109 (2018).

52. Jin, W., Grant, J. R., Stothard, P., Moore, S. S. & Guan, L. L. Characterization of bovine mirnas by sequencing and bioinformatics analysis. *BMC Mol. Biol.* **10**, 90 (2009).

53. Sun, J. *et al.* Identification and profiling of conserved and novel micrornas from chinese qinchuan bovine longissimus thoracis. *BMC Genomics* **14**, 42 (2013).

54. Huang, Y. *et al.* Genome-wide dna methylation profiles and their relationships with mrna and the microrna transcriptome in bovine muscle tissue (bos taurine). *Sci. reports* **4**, 6546 (2014).

55. Sun, J. *et al.* Comparative transcriptome analysis reveals significant differences in microrna expression and their target genes between adipose and muscular tissues in cattle. *PLOS ONE* **9**, 1–9 (2014).

56. Sun, J. *et al.* Altered microrna expression in bovine skeletal muscle with age. *Animal Genet.* **46**, 227–238 (2015).

57. Moisá, S. J., Shike, D. W., Shoup, L. & Loor, J. J. Maternal plane of nutrition during late-gestation and weaning age alter steer calf longissimus muscle adipogenic microrna and target gene expression. *Lipids* **51**, 123–138 (2016).

58. Oliveira, G. B. *et al.* Integrative analysis of micrornas and mrnas revealed regulation of composition and metabolism in nelore cattle. *BMC Genomics* **19**, 126 (2018).

59. Kamli, M. R. *et al.* Expressional studies of the aldehyde oxidase (aox1) gene during myogenic differentiation in c2c12 cells. *Biochem. Biophys. Res. Commun.* **450**, 1291–1296 (2014).

60. Cannon, A. R. *et al.* Palladin expression is a conserved characteristic of the desmoplastic tumor microenvironment and contributes to altered gene expression. *Cytoskelet.* **72**, 402–411 (2015).

61. Jin, L. The actin associated protein palladin in smooth muscle and in the development of diseases of the cardiovasculature and in cancer. *J. Muscle Res. Cell Motil.* **32**, 7–17 (2011).

62. Nguyen, N. & Wang, H. Dual roles of palladin protein in in vitro myogenesis: Inhibition of early induction but promotion of myotube maturation. *PLOS ONE* **10**, e0124762 (2015).

63. Saatchi, M. *et al.* Qtls associated with dry matter intake, metabolic mid-test weight, growth and feed efficiency have little overlap across 4 beef cattle studies. *BMC Genomics* **15**, 1004 (2014).

64. Barendse, W. J. Dna markers for meat tenderness. *Int. patent publication WO 02/064820 A1* (2002).

65. Tait, R. G. *et al.* Capn1, cast, and dgat1 genetic effects on preweaning performance, carcass quality traits, and residual variance of tenderness in a beef cattle population selected for haplotype and allele equalization. *J. Animal Sci.* **92**, 5382–5393 (2014).

66. Joyce, P. I. *et al.* Deficiency of the zinc finger protein zfp106 causes motor and sensory neurodegeneration. *Hum. Mol. Genet.* **25**, 291–307 (2016).

67. Anderson, D. M. *et al.* Severe muscle wasting and denervation in mice lacking the rna-binding protein zfp106. *Proc. Natl. Acad. Sci.* **113**, E4494–E4503 (2016).

68. Celona, B. *et al.* Suppression of *C9orf72* rna repeat-induced neurotoxicity by the als-associated rna-binding protein zfp106. *eLife* **6**, e19032 (2017).

69. Casey, L. M., Lyon, H. D. & Olmsted, J. B. Muscle-specific microtubule-associated protein 4 is expressed early in myogenesis and is not sufficient to induce microtubule reorganization. *Cell Motil.* **54**, 317–336 (2003).

70. Mogessie, B., Roth, D., Rahil, Z. & Straube, A. A novel isoform of map4 organises the paraxial microtubule array required for muscle cell differentiation. *eLife* **4**, e05697 (2015).

71. Venuti, J. M., Morris, J. H., Vivian, J. L., Olson, E. N. & Klein, W. H. Myogenin is required for late but not early aspects of myogenesis during mouse development. *The J. Cell Biol.* **128**, 563–576 (1995).

72. Hasty, P. *et al.* Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene. *Nat.* **364**, 501–506 (1993).

Acknowledgements

The sampling of the Limousin *Longissimus thoraci* biopsies was part of the Qualvigene project, funded by Agence Nationale de la Recherche (contracts ANR05-GANI-005 and ANR-05-GANI-017-01) and APIS-GENE (contract 01-2005-QualviGenA-02). The WGS work was funded by the French National Research Agency (Regulomix project, contract ANR-09-GENM-011). The RNA-Seq work was funded by the INRA Animal Genetics Department (BovRNA-Seq project). We are grateful to the Genotoul bioinformatics facility for providing computing and storage resources. We would also like to thank the ABIES doctoral school for funding G.G. PhD fellowship.

Author contributions

G.G. performed data analyses. D.R. designed the experiments, secured the funding and supervised the project. A.E.H. performed Pyrosequencing experiments. C.M. and E.R. prepared RNA and DNA samples. D.E. supervised the Illumina sequencing. R.L. performed the CNV detection. M.S., N.H. and A.V performed the SNP validation by Sanger sequencing. E.B. and N.B. helped with the Pyrosequencing experiments and analyses. C.J.V.J and A.J.C. contributed the ASE Holstein data. G.G. and D.R. drafted the manuscript and A.J.C. revised it.

Additional information

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

Supplementary data

SNP ID	Forward primer (5′ → 3′)	Reverse primer (5′ → 3′)	Tm, °C	Amplicon length	Sequencing primer (5′ → 3′)
rs382378456	bio-GGCCAACAGATTGTTTCAATT	CCGGTCAATGTGATAAAATCTT	47.8	118	AAAAGTTTTGAGTCTGATAC
rs208775256	ACGGGAAGAGTGAGGGTCA	bio-TGGCCCCCATCAGTGACT	68.2	61	ACGCGCTGCGGAGGTG
rs41255587	bio-TCTCCCCACAGTGCCTGTAAA	ACGTGAGGCATCGTTTTCC	51.1	86	AACAACAAAAATAAGTGCA
rs209641420	bio-GAATTCCTGCTTTCAGAGAGTACA	TGAAATTGCCAAGTCCTAAGAAA	48.6	73	GAAATAGTTTAAACAGATGC
rs110694123	bio-GCCAAGCCCAGAACCAGG	GGCGCTCCTTGGGTCTGA	52.9	96	CAGGTACCCGCTGTC
rs109919583	bio-TAAGCTGATGGGGTCTACGG	GGCTAACCCTTGGTCTTAGTACCT	59.0	60	CCTGGTCCAGCTGGG

Table S1. Primers used for the Pyrosequencing validation.

Animal	Number of reads	Number of bases (in Gb)	Number of mapped reads	% mapped reads	Properly paired reads	% properly paired reads
LIM1	324,087,018	32.73	221,015,607	68.20	184,684,650	56.99
LIM2	341,226,956	34.46	305,598,586	89.56	297,283,336	87.12
LIM3	349,468,660	35.30	295,083,023	84.44	275,228,392	78.76
LIM4	399,160,768	40.32	323,013,864	80.92	305,644,324	76.57
LIM5	258,474,904	26.11	236,719,259	91.58	225,434,012	87.22
LIM6	274,588,874	27.73	232,926,147	84.83	218,564,660	79.60
LIM7	184,667,490	18.65	150,916,707	81.72	136,235,566	73.77
LIM8	330,401,950	33.37	269,954,137	81.70	258,060,878	78.11
LIM9	349,760,660	35.33	319,442,682	91.33	303,101,250	86.66
LIM13	124,730,546	12.60	111,645,197	89.51	50,428,016	40.43
LIM14	266,660,768	26.93	232,007,905	87.00	216,575,916	81.22
LIM15	223,343,850	22.56	180,775,911	80.94	163,669,548	73.28
LIM16	371,989,338	37.57	340,900,866	91.64	330,715,116	88.90
LIM17	92,033,124	9.30	75,520,340	82.06	65,979,448	71.69
LIM18	380,841,540	38.46	352,017,103	92.43	333,859,586	87.66
LIM19	237,734,500	24.01	192,411,974	80.94	183,106,422	77.02
LIM20	304,990,940	30.80	249,443,020	81.79	208,210,852	68.27
LIM21	152,242,676	15.38	130,037,102	85.41	122,791,084	80.65
LIM22	355,455,048	35.90	300,611,692	84.57	288,120,118	81.06
Total	5,321,859,610	537.51	4,477,661,420	84.14	4,163,525,080	78.23

Table S2. Results of WGS read mapping.

Ind	Number of reads	Number of bases (in Gb)	Number of mapped reads	% mapped reads	Number of uniquely mapped reads	% uniquely mapped reads	Number of secondary alignments	Number of non-unique alignments	Aligned to genes	Exonic	Intronic	Intergenic	Intronic/ intergenic overlapping exon	Reference
LIM1	86,352,760	8.635276	82,444,552	95.47	77,485,468	89.73	4,959,084	7,463,794	53,292,640	71.19%	11.22%	17.59%	6.87%	[24]
LIM2	72,251,962	7.2251962	65,725,502	90.97	62,418,616	86.39	3,306,886	5,204,744	42,756,557	70.71%	11.08%	18.21%	6.18%	[24]
LIM3	90,678,870	9.067887	85,827,528	94.65	80,735,158	89.03	5,092,370	8,033,792	55,171,586	71.03%	11.37%	17.60%	6.32%	[24]
LIM4	74,649,210	7.464921	73,105,044	97.93	67,755,422	90.77	5,349,622	7,597,568	47,256,659	72.28%	10.85%	16.88%	6.44%	[25]
LIM5	72,416,218	7.2416218	67,650,068	93.42	63,934,406	88.29	3,715,662	5,893,836	41,172,927	66.72%	12.94%	20.34%	6.94%	[25]
LIM6	80,220,062	8.0220062	76,279,012	95.09	72,100,440	89.88	4,178,572	6,550,162	49,768,744	71.47%	10.95%	17.58%	6.48%	[25]
LIM7	48,943,584	4.8943584	46,535,364	95.08	44,395,770	90.71	2,139,594	3,368,832	31,104,845	72.13%	11.69%	16.18%	7.18%	[25]
LIM8	99,643,662	9.9643662	94,676,802	95.02	89,827,428	90.15	4,849,374	7,772,206	61,977,586	71.44%	11.30%	17.26%	7.34%	[25]
LIM9	59,836,970	5.983697	56,958,018	95.19	53,706,558	89.75	3,251,460	5,068,284	36,961,321	71.31%	10.91%	17.78%	6.24%	[25]
LIM13	84,529,478	8.4529478	80,418,818	95.14	75,504,072	89.32	4,914,746	7,686,154	52,182,535	71.83%	10.51%	17.66%	6.40%	[26]
LIM14	70,268,370	7.026837	63,829,866	90.84	60,602,820	86.24	3,227,046	5,153,988	39,125,296	66.74%	12.20%	21.06%	6.99%	[26]
LIM15	73,757,322	7.3757322	68,270,584	92.56	64,697,736	87.72	3,572,848	5,519,324	44,288,481	70.69%	10.80%	18.51%	6.36%	[26]
LIM16	69,584,672	6.9584672	62,849,226	90.32	59,612,996	85.67	3,236,230	5,179,470	38,180,352	66.27%	12.56%	21.18%	6.93%	[26]
LIM17	50,826,018	5.0826018	48,206,484	94.85	45,684,520	89.88	2,521,964	3,926,020	31,630,075	71.52%	10.75%	17.73%	6.20%	[26]
LIM18	60,787,486	6.0787486	57,283,860	94.24	54,165,520	89.11	3,118,340	4,918,798	36,982,641	70.72%	11.63%	17.66%	6.24%	[26]
LIM19	77,601,664	7.7601664	71,890,464	92.64	68,073,012	87.72	3,817,452	6,048,402	46,706,441	71.02%	11.01%	17.97%	6.55%	[26]
LIM20	34,501,330	3.450133	33,478,554	97.04	31,363,812	90.91	2,114,742	3,090,726	21,862,234	72.06%	10.53%	17.41%	6.47%	[26]
LIM21	54,290,636	5.4290636	53,837,870	99.17	49,242,526	90.70	4,595,344	6,538,258	34,964,006	74.10%	10.45%	15.45%	6.70%	This Study
LIM22	179,545,394	17.954539	177,943,052	99.11	158,297,776	88.17	19,645,276	26,365,002	106,419,650	70.34%	11.56%	18.10%	6.38%	This Study
Average	75,825,561	7.5825561	71,958,456	94.67	67,347,582	88.95	4,610,874	6,914,703	45,884,451	70.71%	11.28%	18.01%	6.59%	

Table S3. Results of RNA-Seq read mapping.

BTA	Position	ID	Ref	Alt	Gene	Transcript	Variant_type	Animals	#Ref	#Alt	#Total	Binom_pval
1	361646	rs381503661	G	A	ENSBTAG00000020035	ENSBTAT00000037243	3_prime_UTR	LIM5, LIM18	106 ; 182	58 ; 106	164 ; 288	2.21E-04 ; 8.84E-06
1	362318	rs110943703	T	C	ENSBTAG00000020035	ENSBTAT00000037243	3_prime_UTR	LIM6	122	216	338	3.58E-07
1	362426	rs377871984	G	A	ENSBTAG00000020035	ENSBTAT00000037243	3_prime_UTR	LIM5, LIM16, LIM4, LIM20, LIM21	322 ; 206 ; 96 ; 133 ; 277	205 ; 123 ; 63 ; 93 ; 219	527 ; 329 ; 159 ; 226 ; 496	3.92E-07 ; 5.53E-06 ; 1.09E-02 ; 9.33E-03 ; 1.04E-02
1	362525	rs109816298	C	T	ENSBTAG00000020035	ENSBTAT00000037243	3_prime_UTR	LIM5, LIM16, LIM20	116 ; 83 ; 86	182 ; 131 ; 135	298 ; 214 ; 221	1.57E-04 ; 1.26E-03 ; 1.19E-03
1	362724	rs385103051	A	C	ENSBTAG00000020035	ENSBTAT00000037243	3_prime_UTR	LIM5, LIM16, LIM15, LIM17, LIM14	88 ; 57 ; 92 ; 242 ; 395	141 ; 93 ; 125 ; 312 ; 317	229 ; 150 ; 217 ; 554 ; 712	5.61E-04 ; 4.11E-03 ; 2.96E-02 ; 3.34E-03 ; 3.87E-03
1	466250	rs133666463	T	C	ENSBTAG00000011528	ENSBTAT00000015319	synonymous	LIM17, LIM6	115 ; 223	85 ; 278	200 ; 501	4.00E-02 ; 1.58E-02
1	466306	rs135162378	A	G	ENSBTAG00000011528	ENSBTAT00000015319	synonymous	LIM21	249	188	437	4.05E-03
1	701152	rs137409022	A	G	ENSBTAG00000012594	ENSBTAT00000016717	intron	LIM8	28	14	42	4.36E-02
1	1376977	.	A	G	ENSBTAG00000012899	ENSBTAT00000017147	synonymous	LIM4, LIM13	112 ; 112	218 ; 173	330 ; 285	5.57E-09 ; 3.62E-04
1	6366794	rs209074791	G	A	ENSBTAG00000007444	ENSBTAT00000049195	3_prime_UTR	LIM19	40	23	63	4.30E-02
1	6366864	rs43215583	C	T	ENSBTAG00000007444	ENSBTAT00000049195	3_prime_UTR	LIM14	16	31	47	4.00E-02
1	6366938	rs137382668	T	G	ENSBTAG00000007444	ENSBTAT00000049195	3_prime_UTR	LIM19, LIM14	31 ; 34	12 ; 17	43 ; 51	5.40E-03 ; 2.41E-02
1	6480486	rs136043966	G	C	ENSBTAG00000014233	ENSBTAT00000018917	missense	LIM18	103	144	247	1.08E-02
1	6481146	rs207697817	A	G	ENSBTAG00000014233	ENSBTAT00000018917	missense	LIM20, LIM2	98 ; 175	69 ; 139	167 ; 314	3.00E-02 ; 4.81E-02
1	6531508	rs383181128	A	G	ENSBTAG00000020121	ENSBTAT00000026800	intron	LIM6	25	12	37	4.70E-02
1	6531985	rs207665682	G	A	ENSBTAG00000020121	ENSBTAT00000026800	intron	LIM3	21	8	29	2.41E-02
1	6532008	rs110269089	C	T	ENSBTAG00000020121	ENSBTAT00000026800	intron	LIM13, LIM6	11 ; 7	24 ; 18	35 ; 25	4.10E-02 ; 4.33E-02
1	6603707	rs378504844	A	G	ENSBTAG00000000201	ENSBTAT00000032432	3_prime_UTR	LIM21	6	20	26	9.36E-03
1	6603985	rs43216926	T	C	ENSBTAG00000000201	ENSBTAT00000032432	downstream_gene	LIM6	9	24	33	1.35E-02

Table S4. Extract of the Table with information on predicted ASE SNPs. REF and ALT are the reference and alternative alleles, respectively. #REF is the number of reads mapped on the reference allele. #ALT is the number of reads mapped on the alternative allele. #Total is the number of reads mapped on this SNPs. Binom pval is the *p*-value of the binomial test computed with #REF and #ALT data. When there are two or more individuals, the results in the four last columns are in the same order than in the Animals column, separated by a semicolon.

Animal	Het RNA SNP	Het DNA SNP	Het SNP for both	ASE-SNP
LIM1	21,618	3,079,656	11,354	634
LIM2	21,791	3,062,240	12,671	696
LIM3	22,683	3,532,660	16,413	921
LIM4	17,434	3,876,950	12,968	991
LIM5	16,231	3,315,943	11,637	728
LIM6	20,307	3,354,198	12,860	654
LIM7	7,020	2,855,384	3,439	184
LIM8	24,489	3,518,227	17,430	747
LIM9	16,172	2,830,312	9,012	508
LIM13	19,683	1,047,724	3,671	239
LIM14	15,142	3,212,066	9,932	606
LIM15	17,954	2,996,669	10,135	619
LIM16	15,964	3,562,784	11,854	712
LIM17	17,049	1,214,141	3,677	207
LIM18	11,985	3,619,926	8,585	480
LIM19	19,812	3,011,647	11,315	674
LIM20	10,681	2,917,059	4,423	280
LIM21	14,753	2,333,694	6,466	383
LIM22	38,043	3,124,434	24,815	652

Table S5. Distribution of ASE-SNPs per individual.

ASE-SNP informations									QTL informations							
BTA	Position	SNP ID	REF	ALT	Gene	Transcript	Variant_type	Animals	BTA	QTL_start	QTL_end	QTL ID	Global_trait	trait	Abbrev	Breed
1	107184976	rs385331844	G	A			intergenic_variant	LIM2	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
1	107213401	rs43260727	T	C	ENSBTAG00000046104	ENSBTAT00000064763	synonymous_variant	LIM2	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
1	107213611	rs43260724	A	G	ENSBTAG00000046104	ENSBTAT00000064763	synonymous_variant	LIM2	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
1	107214222	rs207742155	G	A	ENSBTAG00000046104	ENSBTAT00000064763	downstream_gene_variant	LIM9	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
1	107214245	rs209391400	T	C	ENSBTAG00000046104	ENSBTAT00000064763	downstream_gene_variant	LIM2	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
1	107214252	rs211102958	A	G	ENSBTAG00000046104	ENSBTAT00000064763	downstream_gene_variant	LIM2	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
1	107214346	rs208201426	G	A	ENSBTAG00000046104	ENSBTAT00000064763	downstream_gene_variant	LIM2	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
1	107214445	rs379441420	C	T	ENSBTAG00000046104	ENSBTAT00000064763	downstream_gene_variant	LIM2	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
1	107214735	rs210888907	A	G	ENSBTAG00000046104	ENSBTAT00000064763	downstream_gene_variant	LIM2	1	106448523	107216178	106707	Exterior_QTL	Udder swelling score	USS	limousin
2	129038751	rs379529624	C	G	ENSBTAG00000013048	ENSBTAT00000017344	downstream_gene_variant	LIM19	2	128497489	129296987	106710	Exterior_QTL	Udder swelling score	USS	limousin
3	16052490	rs110040286	A	T	ENSBTAG00000007519	ENSBTAT00000009896	downstream_gene_variant	LIM19	3	15923339	16285883	106711	Exterior_QTL	Udder swelling score	USS	limousin
3	27659167	rs208834893	A	G	ENSBTAG00000011500	ENSBTAT00000015285	missense_variant	LIM8,LIM1	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27703005	rs109040459	T	C	ENSBTAG00000011500	ENSBTAT00000015285	synonymous_variant	LIM8,LIM19	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27704665	rs110118814	T	G	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27704859	rs109035676	C	T	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27704904	rs109830169	C	T	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27704920	rs110948336	C	T	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM1,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705042	rs109741886	A	G	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705057	rs110664626	G	A	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705074	rs110780605	A	G	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705186	rs109194502	G	A	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705223	rs109982146	A	G	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705245	rs109336225	G	C	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705438	rs109090980	G	A	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705576	rs136999734	T	C	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8,LIM1,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705619	rs110952911	T	C	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM1,LIM3,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705635	rs386107473	G	C	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8,LIM1,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705759	rs380253473	G	A	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8,LIM1,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27705953	rs383709488	A	G	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27706132	rs110676946	T	C	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM8,LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27706186	rs109164555	C	T	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM22	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin
3	27706267	rs109216992	T	C	ENSBTAG00000011500	ENSBTAT00000015285	intron_variant	LIM3	3	27387215	28034626	106681	Milk_QTL	Milk yield	MY	limousin

Table S6. Extract of Table with information on ASE SNPs within QTLs.

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b	GO_type
GO:0044449	contractile fiber part	6.41E-18	1.15E-14	3.03	13586	156	1867	65	Component
GO:0044424	intracellular part	3.38E-16	3.04E-13	1.08	13586	10949	1867	1628	Component
GO:0030018	Z disc	1.90E-13	1.14E-10	3.38	13586	86	1867	40	Component
GO:0044444	cytoplasmic part	2.31E-12	1.04E-09	1.13	13586	7594	1867	1181	Component
GO:0043226	organelle	3.41E-12	1.23E-09	1.12	13586	8392	1867	1286	Component
GO:0003779	actin binding	3.33E-13	1.43E-09	2.1	13586	329	1867	95	Function
GO:0043229	intracellular organelle	5.55E-12	1.66E-09	1.14	13586	7256	1867	1133	Component
GO:0043227	membrane-bounded organelle	9.19E-11	2.36E-08	1.12	13586	7584	1867	1169	Component
GO:0030029	actin filament-based process	9.55E-12	1.37E-07	2.12	13586	282	1867	82	Process
GO:0008092	cytoskeletal protein binding	9.09E-11	1.95E-07	1.63	13586	705	1867	158	Function
GO:0006936	muscle contraction	2.71E-10	1.30E-06	2.41	13586	160	1867	53	Process
GO:0003012	muscle system process	2.39E-10	1.72E-06	2.24	13586	201	1867	62	Process
GO:0008307	structural constituent of muscle	4.58E-09	6.55E-06	4.19	13586	33	1867	19	Function
GO:0051015	actin filament binding	6.91E-09	7.40E-06	2.48	13586	123	1867	42	Function
GO:0005829	cytosol	5.29E-08	1.05E-05	1.18	13586	4037	1867	654	Component
GO:0051371	muscle alpha-actinin binding	1.34E-08	1.15E-05	5.82	13586	15	1867	12	Function
GO:0043231	intracellular membrane-bounded organelle	5.17E-08	1.16E-05	1.13	13586	6048	1867	938	Component
GO:0045214	sarcomere organization	4.45E-09	1.59E-05	4.58	13586	27	1867	17	Process
GO:0031430	M band	8.99E-08	1.61E-05	4.16	13586	28	1867	16	Component
GO:0044422	organelle part	1.30E-07	2.13E-05	1.1	13586	7356	1867	1114	Component

Table S7. Top 20 most-enriched GO terms. Enrichment = (b/n)/(B/N). N is the total number of genes. B is the total number of genes associated with a specific GO term. n is the number of genes in the top in the target set. b is the number of genes in the intersection.

BTA	Position	SNP ID	Transcript	Gene	Variant consequence	rho	P-value	Bonferroni
1	66717340	rs109294076	ENSBTAT00000019285	FBXO40	downstream_gene	0.64	0.003312312	0.019873871
1	66718953	rs210664723	ENSBTAT00000019285	FBXO40	downstream_gene	0.74	0.000302294	0.001813764
1	66719507	rs384814038	ENSBTAT00000019285	FBXO40	downstream_gene	0.7	0.000840747	0.005044482
1	66719828	rs209433003	ENSBTAT00000019285	FBXO40	downstream_gene	0.77	0.000122733	0.0007364
1	66721006	rs209699784	ENSBTAT00000019285	FBXO40	downstream_gene	0.63	0.004122368	0.02473421
1	66723680	rs110013999	ENSBTAT00000019285	FBXO40	intron	0.74	0.000321862	0.001931174
1	107213611	rs43260724	ENSBTAT00000064763	B3GALNT1	synonymous	0.62	0.004420025	0.039780225
1	107214222	rs207742155	ENSBTAT00000064763	B3GALNT1	downstream_gene	0.75	0.000246573	0.002219161
1	107214245	rs209391400	ENSBTAT00000064763	B3GALNT1	downstream_gene	0.67	0.001715958	0.015443626
1	107214252	rs211102958	ENSBTAT00000064763	B3GALNT1	downstream_gene	0.67	0.001715958	0.015443626
1	107214346	rs208201426	ENSBTAT00000064763	B3GALNT1	downstream_gene	0.67	0.001715958	0.015443626
1	107214735	rs210888907	ENSBTAT00000064763	B3GALNT1	downstream_gene	0.62	0.00444923	0.04004307
1	107214967	rs43260723	ENSBTAT00000064763	B3GALNT1	downstream_gene	0.62	0.00444923	0.04004307
1	112650588	rs134837808	ENSBTAT00000017301	GMPS	intron	-0.6	0.006501513	0.006501513
1	151133143	rs132945750	ENSBTAT00000050511	TTC3	intron	-0.6	0.006472771	0.032363857
1	151139707	rs133423051	ENSBTAT00000050511	TTC3	3_prime_UTR	-0.68	0.001279486	0.006397431
1	151139997	rs209579998	ENSBTAT00000050511	TTC3	3_prime_UTR	-0.68	0.001279486	0.006397431
1	151140026	rs136904677	ENSBTAT00000050511	TTC3	3_prime_UTR	-0.68	0.001279486	0.006397431
1	151140147	rs135034351	ENSBTAT00000050511	TTC3	3_prime_UTR	-0.6	0.006472771	0.032363857

Table S8. Extract of Table of ASE SNPs with a significant correlation between transcript expression and SNP genotype.

Gene	SNP	ALT	REF	Total
ANKRD16	rs137842237	0	0	0
CCDC90B	rs136608384	14	15	29
FAM32A	rs41255136	32	29	61
FAM32A	rs43498240	22	36	58
IMP3	rs133150935	39	37	76
IMP3	rs134305019	23	46	69
IMP3	rs137844508	23	20	43
PRNP	rs109763272	5	9	14
TTC3	rs133423051	0	0	0
TTC3	rs135034351	60	14	74
TTC3	rs136904677	21	16	37
TTC3	rs209579998	30	43	73
TYK2	rs109497730	0	0	0
TYK2	rs208406423	34	6	40

Table S9. Detection of miRNA binding sites impacted by SNPs in 3'UTR regions. REF is the number of miRNAs binding only SNP with reference allele. ALT is the number of miRNA binding only SNP with alternative allele. Total is the number of miRNAs bond impacted by the SNP.

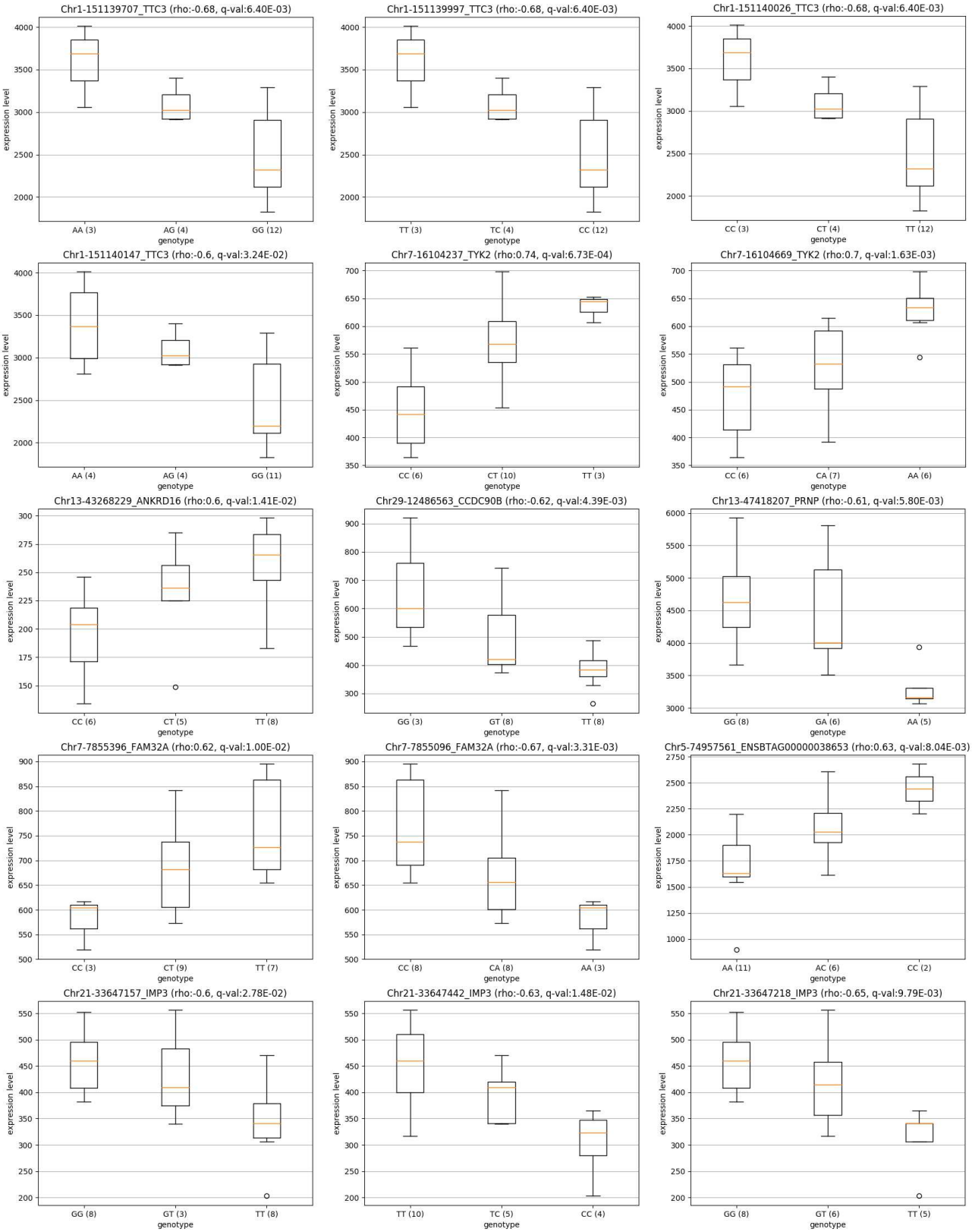

Figure S1. Boxplots of significantly correlated SNPs in 3'UTR regions. (N) number of animals per genotype.

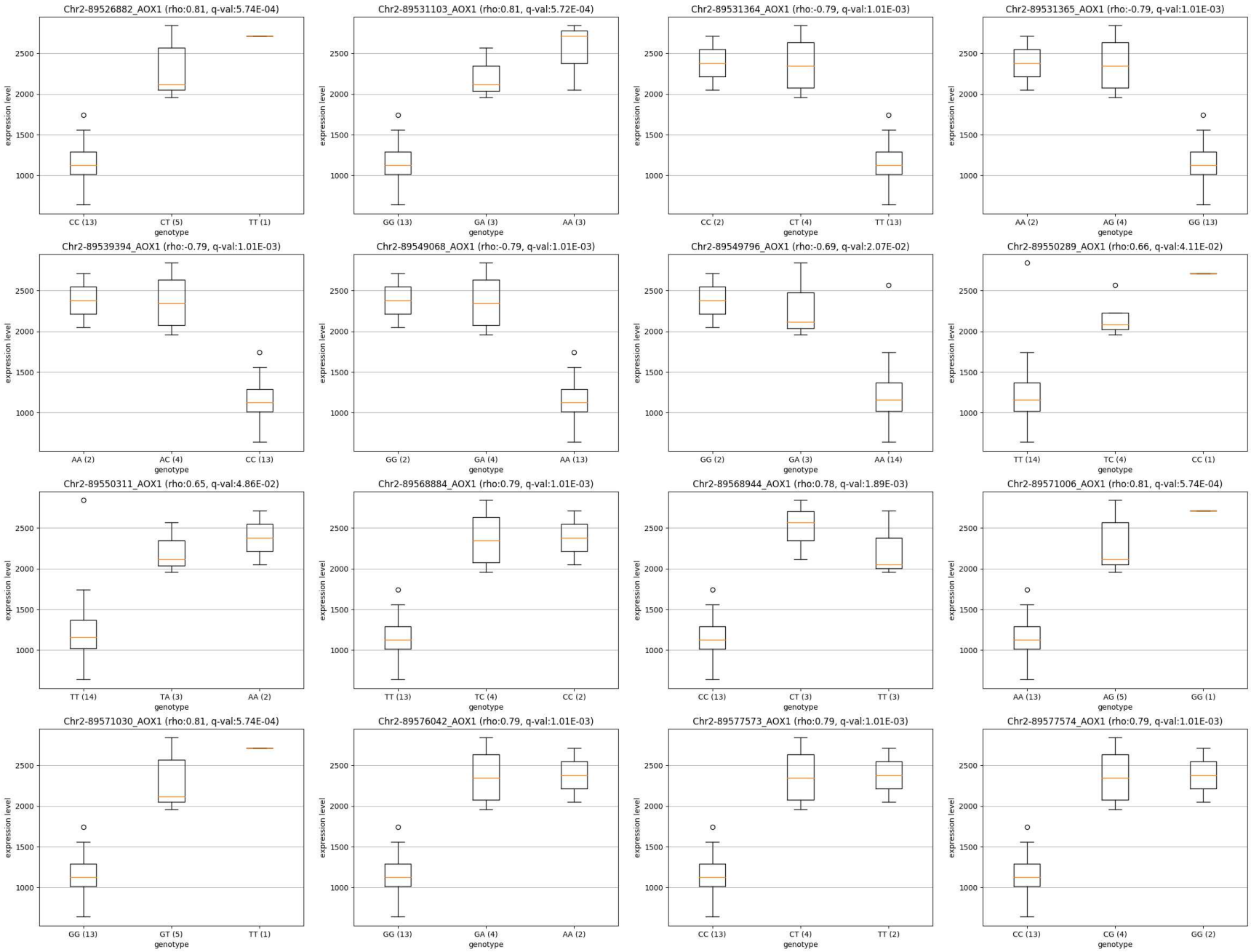

Figure S2. Boxplots of significantly correlated SNPs within *AOX1*. (N) number of animals per genotype.

Article 2 : Etude de l'expression allèle-spécifique intra individu et inter tissu chez la vache holstein.

Article en préparation.

A. Introduction

La régulation des gènes est un processus fondamental dans le développement et la maintenance de l'organisme. Chez les mammifères, la variabilité de l'expression des gènes est un phénomène courant (Segal *et al.*, 2003; Amit *et al.*, 2009). C'est pourquoi il est important d'étudier cette variabilité afin de comprendre la régulation des gènes. Différentes approches existent dans ce but : les analyses de locus de caractères quantitatifs associé à l'expression (eQTL) et d'expression allèle-spécifique (ASE). La combinaison des deux approches est très efficace pour détecter les régulations de l'expression des gènes en *trans* et surtout en *cis*.

Un eQTL est une région d'ADN avec quelques différences de séquences nucléotidiques (des SNPs, des insertions ou des délétions) qui affectent le niveau d'expression d'un gène en *cis* ou en *trans*. Ils peuvent être identifiés par des études d'association pangénomique de l'expression (eGWAS), une méthode d'analyse qui consiste à calculer la probabilité qu'un polymorphisme affecte l'expression des gènes. Hélas, ce type d'analyse nécessite un grand nombre d'échantillons pour minimiser le taux de faux positifs (Haley et De Koning, 2006).

Plusieurs études d'analyses eQTL ont été menées chez l'homme (Zou *et al.*, 2012; Sabbagh *et al.*, 2016; Grigoryev *et al.*, 2015), ainsi que dans le cadre du projet GTEx (The GTEx Consortium, 2015). Jusqu'à présent chez le bovin, seulement deux ont été publiés dont une très récemment : Lopdell *et al.* (2017) sur des vaches laitières de race Frisonne, Jersiaise et mixte (Jersiaise x Frisonne) pour étudier les SNPs liés à la composition du lait en lactose et Higgins *et al.* (2018) sur des bovins irlandais (principalement des hybrides de race Charolaise, Limousine, Bleu-Belge, Simmental et Angus) pour étudier les SNPs liés à l'efficacité alimentaire.

L'analyse ASE est une approche robuste pour quantifier la variation de l'expression entre deux haplotypes d'un individu pour des sites hétérozygotes (Castel *et al.*, 2015). Cette approche est complémentaire aux analyses eQTL pour identifier les variants qui altèrent l'expression des gènes

grâce à un meilleur contrôle intra-individuel qui élimine ainsi les influences génétique en *trans* et les facteurs environnementaux (Pastinen, 2010). On peut aussi utiliser un nombre d'échantillons plus faible (Muráni *et al.*, 2009).

Des études pangénomiques ASE ont été réalisées dans différentes espèces comme l'humain (Chen *et al.*, 2016), la souris (Lagarrigue *et al.*, 2013) ou la mouche (Fear *et al.*, 2016) ainsi que dans différentes espèces d'élevage : porc (Maroilley *et al.*, 2017), poulet (Zhuo *et al.*, 2017) ou mouton (Ghazanfar *et al.*, 2017). De plus, certains gènes ASE ont été identifiés comme impactant des caractères d'importance économique (Esteve-Codina *et al.*, 2011; Muráni *et al.*, 2009). Chez le bovin, quelques études ont été réalisés en holstein (Chitwood *et al.*, 2013; Chamberlain *et al.*, 2015). Chitwood *et al.* (2013) ont découvert dans leur étude 473 ASE-SNPs à partir de 5 blastocytes bovins (parmi 2 524 SNPs hétérozygotes différents). Chamberlain *et al.* (2015) ont prédit dans leur étude 19 082 ASE-SNPs (1 060 en moyenne par tissu) à partir de 18 tissus différents provenant d'une seule vache laitière.

Dans cette étude, nous avons effectué une recherche de l'expression allèle-spécifique à l'aide d'échantillons de 8 tissus provenant de 6 vaches laitières de race Holstein. Pour réaliser cette étude, nous avons utilisé les séquences du génome entier et les données RNA-Seq de tous les tissus prélevés. Cette étude permet d'approfondir les précédentes analyses ASE effectuées chez la Holstein. De plus, nous avons tenté de déterminer le lien entre le déséquilibre allélique et une expression tissu-spécifique.

B. Matériels et méthodes

B.1. Animaux et échantillonnage

Six vaches holstein ont été sélectionnées dans le cadre du projet BOVREG. Elles ont été abattues de manière éthique à l'âge de 26 mois pour la plus jeune et 61 mois pour la plus vieille et elles ont toutes donné naissance à au moins un veau. Huit tissus ont été disséqués immédiatement après la mort : muscle, foie, poumon, rein, rate, cœur, utérus et ovaire. Les échantillons de tissus ont été surgelés dans de l'azote liquide et stockés à $-80\,°C$. Les vaches ont été abattues par un abattoir certifié en accord avec les réglementations française de protection animale (Code Rural, Articles R214-64 à R214-71 ; Legifrance, 2011).

B.2. Séquençage génome entier (WGS) et alignement de séquence

L'ADN a été extrait des 48 échantillons (8 tissus des 6 individus) en utilisant le kit Wizard Genomic DNA Purification (Promega). Chaque échantillon d'ADN purifié a été évalué par électrophorèse sur gel d'agarose. La concentration d'ADN a été mesuré avec un Nanodrop ND-100 (Thermo Fi-

sher Scientific). Les librairies de séquençage ont été préparées en utilisant le kit TruSeq SBS v3-HS (Illumina) et le génome entier a été séquencé en utilisant une approche paired-end de 2×100 pb avec un Illumina HiSeq 2000. L'alignement des séquences a été réalisé avec l'outil Burrows-Wheeler Alignment tool (BWA-v0.6.1-r104 Li *et al.*, 2009) en utilisant les paramètres par défaut pour aligner les lectures sur le génome de référence bovin UMD3.1 (Zimin *et al.*, 2009). Les duplicats de PCR potentiel ont été retirés à l'aide de l'outil MarkDuplicates du logiciel Picard version 1.4.0. Seules les lectures pairées alignées proprement avec une qualité d'alignement d'au moins 30 ($-q = 30$) ont été conservées. Les fichiers BAM ainsi obtenus ont été utilisés pour toutes les analyses suivantes.

B.3. RNA-Seq et alignement de séquence

Après un transfert dans un tampon de lyse glacé RNeasy RLT (Qiagen), les échantillons de tissus ont été homogénisés en utilisant l'homogénéisateur Precellys Evolution (Bertin Technologie). L'ARN total a été isolé à l'aide de colonnes RNeasy Midi (Qiagen) puis traité avec de l'ADNase I exempt de ARNase (Qiagen) durant 15 minutes à température ambiante en accord avec le protocole du fabricant. La concentration de l'ARN total a été mesuré avec un Nanodrop ND-100 (Thermo Scientific) et la qualité a été évaluée avec un kit RNA 6000 Nano Labchip à l'aide d'un Bioanalyseur Agilent 2100 (Agilent Technologies). 46 échantillons ont une valeur de RIN (RNA integrity number) supérieur à 8, les 2 échantillons restants (rein des individus HOL4 et HOL6) ayant une valeur plus faible n'ont pas été retenus pour la suite de cette étude.

Les librairies de RNA-Seq ont été préparées en utilisant le kit TruSeq RNA Sample Preparation (Illumina) en suivant les instructions du fabricant. Brièvement, les molécules d'ARNm avec leur queue polyA ont été purifiées à partir de 4 μg de l'ARN total de chaque échantillon en utilisant des billes magnétiques oligo(dT), puis fragmentées en morceaux de 150 à 400 pb à l'aide de cations divalents à 94 °C pendant 8 minutes. Les fragments d'ARNm clivés ont été convertis en ADN complémentaire (ADNc) double-brin par la reverse transcriptase SuperScript II (Life Technologies) et amorcés avec des primers aléatoires. L'ADNc résultant a été purifié grâce aux billes Agencourt AMPure®XP (Beckman Coulter). Puis, l'ADNc a été soumis à une réparation des extrémités et une phosphorylation pour être à nouveau purifié avec les billes Agencourt AMPure®XP (Beckman Coulter). Ces fragments d'ADNc réparés ont été adénylés en 3' produisant des fragments d'ADNc avec une simple base 'A' en excédent à l'extrémité 3' pour permettre la ligation d'adaptateurs. Des adaptateurs Illumina contenant des marques d'indexation ont été liés à l'extrémité 3' des fragments d'ADNc suivis par deux phases de purification avec les billes Agencourt AMPure®XP (Beckman

Coulter). Dix tours d'amplification par PCR ont été réalisés par complémentarités des pimers aux extrémités des adaptateurs dans le but d'enrichir la librairie d'ADNc avec adaptateurs. Les produits PCR ont été encore une fois purifiés avec ses billes Agencourt AMPure®XP (Beckman Coulter) et sélectionnés en fonction de leur taille (200 ± 25 pb) sur un gel d'électrophorèse d'agarose Invitrogen (Thermo Scientific). Les librairies ont été verifiées sur un bioanalyseur Agilent Technologies 2100 avec le kit Agilent High Sensitivity DNA et quantifiées par une PCR quantitative avec le kit QPCR NGS Library Quantification (Agilent Technologies). Après quantification, trois librairies différentes d'ADNc marqués ont été regroupées à un ratio égal et une dernière vérification par qPCR a été réalisée après le regroupement. Chaque groupe de librairie a été utilisé pour un séquençage paired-end 2×100 pb sur une lane du séquenceur Illumina HiSeq2000 avec le kit TruSeq SBS v3-HS (Illumina). Après séquençage, les échantillons ont été démultiplexés et les séquences indexées des adaptateurs ont été retirées à l'aide du logiciel CASAVA v1.8.2 d'Illumina. La qualité des séquences brutes des lectures a été évaluée avec FastQC et Qualimap (Okonechnikov *et al.*, 2016).

La séquence du génome de référence de *Bos taurus* a été téléchargée sur Ensembl (version 91, Bos_taurus.UMD3.1.dna.toplevel.fa). Les lectures ont été alignées sur le génome de référence bovin avec l'outil STAR (version 2.4.2a - Dobin *et al.*, 2013). Les paramètres par défaut ont été utilisés pour l'alignement à l'exception de l'alignement sur les introns (*alignIntronMin* : 20 and *alignIntronMax* : 500 000). Les lectures de chaque échantillon ont été alignées séparément sur le génome de référence et seulement les lectures pairées ont été gardées pour l'alignement. Le nombre de lectures pairées s'alignant de manière unique sur les régions transcrites de chaque transcrit a été calculé pour tous les gènes du transcriptome annoté.

B.4. Identification et annotation des SNPs

Les SNPs ont été détectés suivant les *best practices* avec l'outil HaplotypeCaller de GATK (version 3.4-46) pour les séquences d'ADN et d'ARN respectivement (McKenna *et al.*, 2010; Auwera *et al.*, 2013). Les lectures ont été soumises à un réalignement local, puis à un tri des coordonnées, une recalibration des scores de qualité des bases et suppression des duplicats de PCR (seulement pour les données d'ADN). Enfin la détection des SNP et des indels et leur génotypage a été effectuée et les indels ont été réalignées pour retirer les SNPs présents dans des indels. Avec les analyses de GATK, on a utilisé un seuil de Q30 pour le score de confiance minimum avec les paramètres par défaut. Afin de distinguer un génotype homozygote de référence d'un génotype manquant parmi les échantillons analysés, des détections de variants multi-échantillon ont été effectuées. Ces SNPs et

indels ont été annotés à l'aide de VEP (Variant Effect Predictor - McLaren *et al.*, 2016)

B.5. Détection des ASE-SNPs

ASEReadCounter (Castel *et al.*, 2015) a été utilisé pour compter le nombre de lectures par allèle pour chaque SNP. Pour chaque SNP détecté, le nucléotide à la même position dans la séquence du génome de référence bovin a été remplacé par un 'N' afin d'éliminer des biais d'alignement. Seulement les SNPS à la fois hétérozygotes dans les données ARN et ADN ont été conservés afin de retirer les SNPs discordants, potentiellement dû à une empreinte parentale ou de l'édition d'ARN. Seuls les candidats avec au moins 10 lectures alignées pour un des deux allèles ont été conservés. Pour déterminer si le déséquilibre est significatif, un programme *ad hoc* en Python a fait un test binomial contre un ratio allélique de 0,5 avec une *p*-value de 5%.

B.6. Analyse de l'expression tissu-spécifique

Le package python HTSeq-count (Anders *et al.*, 2015) a été utilisé pour compter le nombres de lectures pour chaque transcrit à partir des alignements des données RNA-Seq avec STAR. Le comptage a été normalisé avec le logiciel DESeq2 (Love *et al.*, 2014) et le clustering hiérarchique a été réalisé avec le même outil. Une matrice booléenne des gènes contenant des ASE-SNPs (1) ou non (0) pour chaque échantillon (tissu-individu) a été créée et seuls les gènes possédant au moins un ASE-SNPs parmi les 46 échantillons ont été conservés. Le clustering hiérarchique de cette matrice a aussi été réalisé avec DESeq2.

C. Résultats et discussions

C.1. Statistiques sur les données de séquençage ARN et ADN

Le séquençage des 6 génomes entiers a généré un total de 2,47 milliards de lectures brutes paired-end correspondant à 249,3 Gb. Approximativement, 334 à 479 millions de lectures paired-end ont été obtenus pour chaque individu. En moyenne, 94,4% (93,8% - 94,9%) des lectures paired-end ont été proprement alignées avec BWA sur le génome de référence bovin UMD3.1 (Table 5).

Le séquençage des 46 librairies RNA-Seq a généré un total de 8,1 mille milliards de lectures brutes paired-end correspondant à 822 Gb. Approximativement, 98,6 à 300 millions de lectures paired-end ont été obtenus pour chaque librairie. En moyenne, 92,7% (47,9%-96,4%) des lectures paired-end ont été correctement alignées avec STAR sur le génome de référence bovin UMD3.1 (Table). La

Ind	Nombre de lectures	Nombre de bases (en Gb)	Nombre de lectures alignées	% lectures alignées	Lectures proprement pairées	% lectures proprement pairées	Couverture
HOL1	439 370 548	44,38	427 310 974	97,3%	414 102 392	94,2%	16,04
HOL2	478 726 950	48,35	465 771 333	97,3%	453 183 130	94,7%	17,46
HOL3	334 383 304	33,77	323 398 395	96,7%	313 804 334	93,8%	12,19
HOL4	404 430 544	40,85	392 017 917	96,9%	381 030 655	94,2%	14,74
HOL5	394 467 208	39,84	383 325 818	97,2%	372 837 757	94,5%	14,39
HOL6	416 957 390	42,11	405 589 814	97,3%	395 544 525	94,9%	15,28

TABLE 5 – Résultats de l'alignement des lectures WGS

performance de l'alignement est comparable à d'autres études réalisées chez le bovin avec STAR et le génome de référence UMD3.1. Par exemple, 90% des transcrits des leucocytes ont été alignés chez la frisonne par McLoughlin *et al.* (2014) et 89% des transcrits musculaire lors de notre première étude chez les taurillons limousins.

C.2. Détection des variants

11 973 098 et 1 910 067 SNPs ont été identifiés à partir des données WGS et RNA-Seq respectivement. 11 788 332 $\pm$ 0,24% et 488 469 $\pm$ 28,2% SNPs ont été identifiés en moyenne par individu et 35% et 30,4% des SNPs détectés sont hétérozygotes en moyenne à partir des données WGS et RNA-Seq respectivement.

Parmi les variants détectés à partir des données WGS (Table 6), 8 153 830 (68,10%) ont été identifiés comme intergéniques, 2 870 461 (23,97%) comme intronique, 429 405 (3,59%) comme étant en amont du gène et 412 052 (3,44%) comme étant en aval du gène. 72 721 variants (0,61%) ont été détectés comme étant exonique (53,03% de synonymes, 46,78% de faux-sens et 0,19% de variant modifiant la séquence codante).

Parmi les 1 910 067 variants trouvés à partir des données RNA-Seq (Table 6), 882 599 (46,21%) ont été identifiés comme intronique, 656 535 (34,37%) comme intergénique, 174 431 (9,13%) comme étant en aval du gène et 104 467 (5,47%) comme étant en amont du gène. 59 301 variants (3,1%) ont été détectés comme étant exonique (56,91% de synonymes, 42,98% de faux-sens et 0,11% de variant modifiant la séquence codante).

Les 68,1 % de SNPs identifiés comme intergéniques avec les données WGS sont en accord avec les

	ADN		ARN	
Annotation VEP	Nombre de SNPs	%	Nombre de SNPs	%
intergenic_variant	8 153 830	68,10	656,535	34,37
intron_variant	2 870 461	23,97	882,599	46,21
upstream_gene_variant	429 405	3,59	104,467	5,47
downstream_gene_variant	412 052	3,44	174,431	9,13
synonymous_variant	38 565	0,32	33,747	1,77
missense_variant	34 021	0,28	25,487	1,33
3_prime_UTR_variant	18 887	0,16	18,203	0,95
splice_region_variant	6 353	0,05	5,029	0,26
5_prime_UTR_variant	4 176	0,03	3,375	0,18
non_coding_transcript_exon_variant	4 110	0,03	4,156	0,22
Unindentified	442	0,00	170	0,01
splice_donor_variant	316	0,00	1,163	0,06
splice_acceptor_variant	221	0,00	578	0,03
coding_sequence_variant	135	0,00	67	0,00
mature_miRNA_variant	77	0,00	11	0,00
non_coding_transcript_variant	27	0,00	29	0,00
stop_retained_variant	20	0,00	20	0,00

TABLE 6 – Résumé des annotations VEP des SNPs détectés à partir des données ADN (WGS) et ARN (RNA-Seq).

70,4% de régions du génome bovin étant intergéniques. Cette proportion est aussi similaire à d'autres études chez le bovin. Par exemple, Choi *et al.* (2015) ont détecté 73% de SNPs intergéniques, 26,2% d'introniques, 4,26% de variants en aval du gène et 4,14% de variants en amont du gène. Dans la première étude, on a aussi trouvé des proportions similaires.

Il est intéressant de noter que 34,37% des SNPs identifiés à partir des données RNA-Seq sont intergéniques. Il s'agit probablement de SNPs présents dans des transcrits de lncRNAs. Dans le cas des 34,37% des SNPs identifiées comme étant introniques, ils ont probablement été détectés à partir des ARN pré-mature (avant l'épissage).

C.3. Comparaison des SNPs

Les SNPs détectés à partir de données WGS et RNA-Seq ont été comparés pour chaque individu et pour chaque tissu. Parmi les SNPs détectés, en moyenne, 11 306 306 sont spécifiques aux données ADN (sur 11 788 332 SNPs détectés en moyenne par individu), 40 112 sont spécifiques aux données ARN (sur 488 469 SNPs détectés par échantillon) et 75 362 pour les SNPs communs aux deux types de données. Pour l'étude, seulement les SNPs en commun sont conservés et la concordance des génotypes ARN et ADN de ces SNPs a été vérifiée pour chaque échantillon. En moyenne, 91,5% (53,1% à 95,8%) des SNPs sont concordants (54,7% strictement homozygotes et 45,3% strictement hétérozygotes) et les 8,5% restants sont donc discordants. En moyenne, 58% des SNPs discordants sont homozygotes pour les séquences d'ADN et hétérozygotes pour les séquences d'ARN, cette discordance peut être en partie expliquée par l'édition d'ARN. 42% des SNPs discordants sont à l'inverse hétérozygotes pour les séquences d'ADN et homozygotes pour les séquences d'ARN, cela peut être expliqué par une expression mono-allélique, notamment avec le phénomène d'empreinte parentale. Enfin, cette différence entre les génotypes ADN et ARN peut être due à des erreurs de séquençage. Pour l'étude du déséquilibre allélique, seulement les SNPs avec un génotype concordant et strictement hétérozygote sont conservés.

C.4. Identification des ASE-SNPs

33 534 ASE-SNPs ont été identifiés parmi les 46 échantillons, 23 962 de ces ASE-SNPS sont spécifiques à un seul échantillon et en moyenne 5 322 ASE-SNPs sont spécifiques à un seul tissu et 4 307 sont spécifiques à un seul individu (Table 7). En moyenne, 1 125 ASE-SNPs ont été détectés par échantillon (min : 438 et max : 2 160) correspondant à 4,81% des SNPs hétérozygotes concordants, toutefois le nombre d'ASE-SNPs détectés semble dépendant du tissu. En effet, en moyenne on trouve environ deux fois plus d'ASE-SNPs dans les échantillons de foie, rate et cœur (respectivement 1 477, 1 460 et 1 348) que dans les échantillons de rein et de muscle (respectivement 640 et 653).

C.5. Expression tissu-spécifique

On constate que 81,2% (27 231 / 33 534, Table 7) des ASE-SNPS détectés sont spécifiques à un seul tissu parmi les 8 étudiés. Afin de comprendre l'influence du choix du tissu sur l'expression allèle-spécifique, on a étudié les différences intra individu inter tissu. Cette différence peut être le résultat de 3 phénomènes :

	HOL1	HOL2	HOL3	HOL4	HOL5	HOL6	Moyenne	Par tissu	Tissu-spécifique
Cœur	1250	1660	975	1184	1799	1219	1348	6449	4199
Foie	2160	1449	910	1138	1827	1380	1477	7262	5530
Muscle	773	438	405	470	1047	785	653	3369	2239
Ovaire	1580	993	863	1221	1159	1186	1167	5771	3456
Poumon	902	1381	906	983	1162	645	997	4949	2763
Rate	1253	1921	1290	1172	1194	1927	1460	6977	4467
Rein	692	511	559		798		640	2302	1339
Utérus	826	1579	645	1214	1009	1317	1098	5499	3238
Moyenne	1180	1242	819	1055	1249	1208	1125	5322	3404
Par individu	8039	8284	5564	6484	8457	7199	7338	33 534	27 231
Individu-spécifique	4784	4792	3035	3865	5184	4183	4307	25 843	23 962

TABLE 7 – Distribution des ASE-SNPs détectés chez les vaches holstein.

1. une expression tissu-spécifique du gène, si ce gène n'est exprimé que dans un seul des tissus étudiés ;
2. une action tissu-spécifique d'un facteur de transcription pouvant entraîner un déséquilibre allélique ;
3. une empreinte parentale (donc intra individu) qui entraîne une expression ASE pour un tissu donné mais qui provoque une expression monoallélique dans les autres tissus.

Tout d'abord, j'ai calculé le niveau d'expression des transcrits et créé un clustering hiérarchique après normalisation sur les 46 échantillons (Figure 24). On constate que les tissus sont regroupés entre eux à l'exception des échantillons de muscle de l'individu HOL2 qui est regroupé avec les échantillons de cœur. De manière attendue, les échantillons de muscle sont proches des échantillons de cœur (le cœur étant un muscle particulier) et les échantillons d'ovaire sont proches de ceux de l'utérus. Les échantillons de poumon et ceux de rate sont regroupés et sont proches du groupe Ovaire-Utérus. Le groupe le plus éloigné des autres tissus est celui qui rapproche les échantillons de rein et de foie. Ces résultats sont similaires avec ceux de Chamberlain *et al.* (2015) hormis deux exceptions : leur étude ne comporte pas de données sur l'utérus et le rein n'est pas groupé avec le foie.

Afin de voir le lien entre les différents échantillons et l'expression allèle-spécifique, nous avons calculé un clustering hiérarchique des échantillons en fonction de la présence ou non d'ASE-SNPS pour chacun des gènes (Figure 25). Contrairement au clustering précédent on ne distingue aucun groupe clair, à part un regroupement avec 3 échantillons de muscles (Muscle-HOL4, Muscle-HOL2 et Muscle-HOL3) et 2 échantillons de rein (Kidney-HOL2, Kidney-HOL3) ainsi qu'un autre regrou-

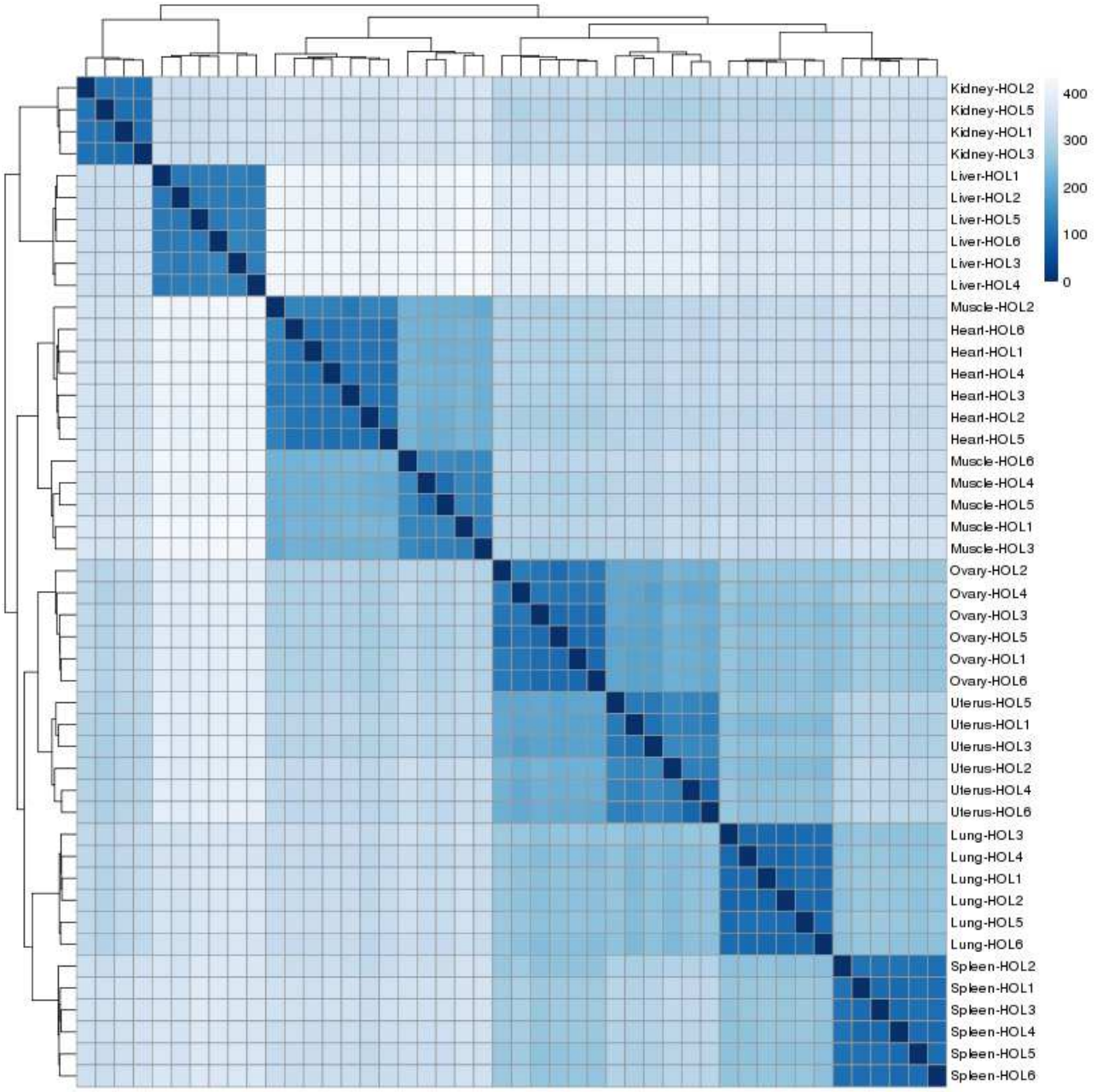

FIGURE 24 – Clustering hiérarchique et heatmap des échantillons (individu-tissu) à partir des données d'expression des gènes. Le niveau de couleur indique la similarité des motifs d'expression entre deux échantillons et la variabilité de l'expression des gènes entre les échantillons est représentée par la hauteur des branches des dendrogrammes.

pement comprenant 4 tissus de l'individu HOL3 (cœur, Poumon, Ovaire et Utérus). Ce qui confirme bien notre intuition de départ sur une expression allèle-spécifique propre à un seul tissu.

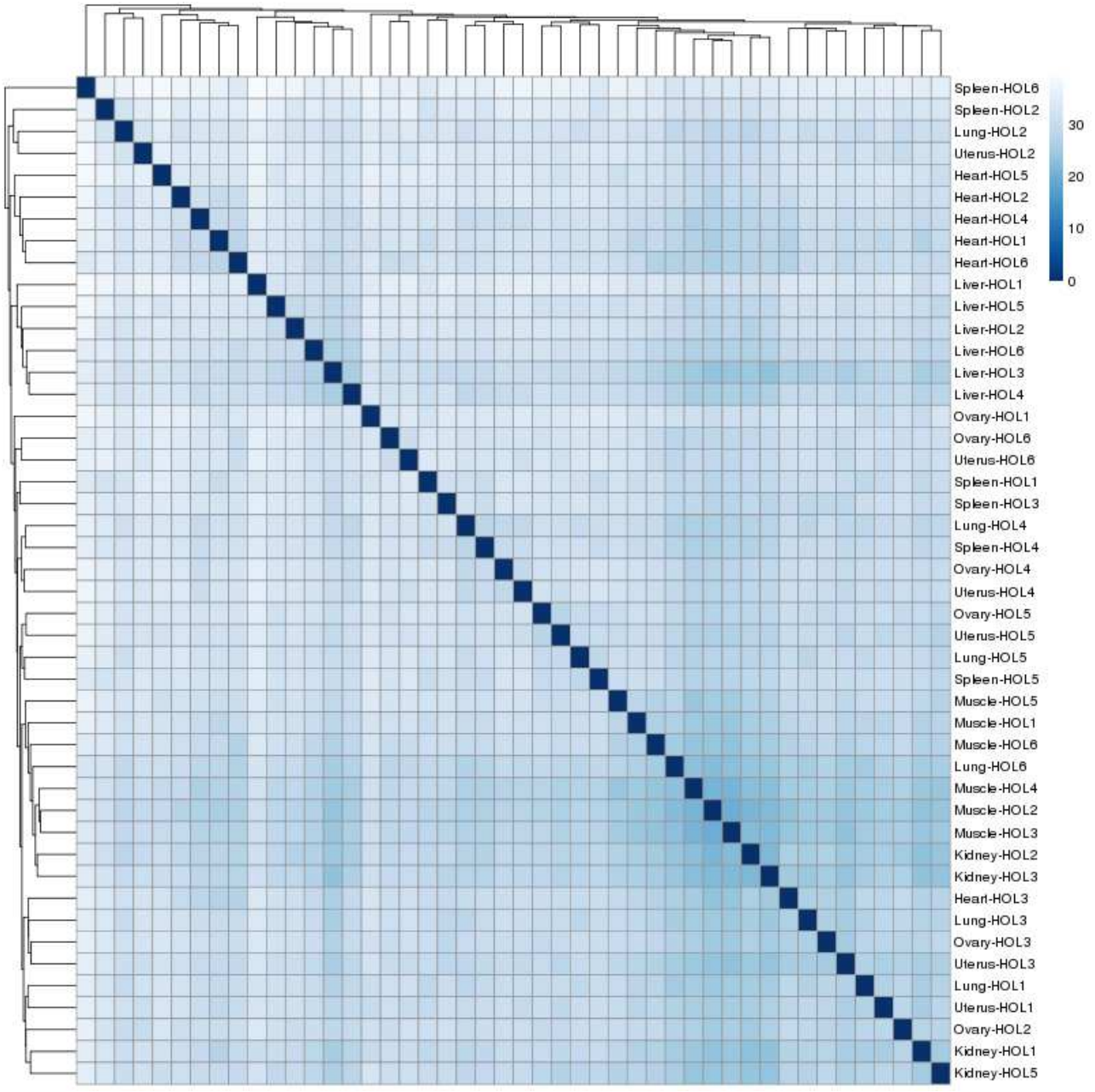

FIGURE 25 – Clustering hiérarchique et heatmap des échantillons (individu-tissu) à partir de la présence d'ASE-SNPs par gène. Le niveau de couleur indique la similarité des motifs d'expression entre deux échantillons et la variabilité entre les échantillons est représentée par la hauteur des branches des dendrogrammes.

Discussion générale

A. Apports des résultats

La première étude a permis de recenser un nombre important de SNPs présents chez la race limousine et d'en identifier certains comme des ASE-SNP. Chez l'homme, de nombreuses études ont été effectuées, notamment par le consortium GTEx (www.gtexportal.org) mais la cartographie des ASE-SNPs est très faible chez le bovin avec seulement une étude à partir d'une seule vache holstein (Chamberlain *et al.*, 2015). Dans l'étude de Chamberlain *et al.* (2015), ils ont en revanche détectés les ASE-SNPs dans 18 tissus différents mais nous avons préféré nous concentrer sur le muscle dans notre étude afin d'étudier des caractères d'intérêt pour les bovins allaitants, notamment ceux associés à la qualité des viandes et à la carcasse.

Avec la deuxième étude, on a pu se focaliser sur plusieurs tissus différents, ce qui nous as permis de constater que certains ASE-SNPs étaient majoritairement spécifiques à un seul tissu. Parmi les restants, on retrouve des ASE-SNPs seulement dans des tissus fonctionnellement proches (cœur et muscle ou utérus et ovaire, par exemple). La répartition des ASE-SNPs parmi les différents tissus et individus est détaillée dans la Table 7 (page 94).

B. Retour sur les approches employées

B.1. Validation à approfondir

Pour la première étude, nous avons testé expérimentalement des ASE-SNPs avec la méthode de pyroséquençage. En raison du coût de la technique (5 € par mesure), nous n'avons pu valider que cinq ASE-SNPs prédits (sur six testés) ce qui est très faible. L'optimal serait de tester expérimentalement tous les ASE-SNPs afin de définir le taux de faux positif du pipeline développé et même de tester tous les SNPs détectés pour définir le taux de faux négatif. Bien entendu, pouvoir tester tous les SNPs et même tous les ASE-SNPs est totalement utopique, en revanche on peut estimer ce taux en choisissant aléatoirement 5% à 10% des ASE-SNPs prédits pour vérifier leur déséquilibre allélique.

Un autre défaut est dans la méthode de pyroséquençage elle-même, en effet elle a une marge d'erreur de 5% (Kreutz *et al.*, 2013) ce qui complique la distinction entre une expression mono-allélique ou un fort déséquilibre d'expression bi-allélique (ratio 95 :5 à 99 :1) ou entre une absence de déséquilibre et une légère expression allèle-spécifique (ratio 45 :55 environ). Deux stratégies sont envisageables pour pallier cette marge d'erreur, soit ajouter un filtre dans le pipeline sur les ratios

problématiques, soit d'utiliser une autre méthode de validation.

B.2. Approche ASE

L'approche ASE à partir des données de RNA-Seq permet de détecter des marqueurs génétiques représentatifs d'un effet *cis*. Les ASE-SNPs sont certainement en déséquilibre de liaison avec le variant causal responsable du contrôle génétique *cis* de l'expression du gène. La difficulté réside à trouver quel est le variant impliqué et quel gène il régule, sachant que plus un SNP est éloigné plus le déséquilibre de liaison (LD) sera faible et plus l'association marqueur-gène régulé sera difficile à déterminer. Les SNP ont été détectés directement dans les données de RNA-Seq et la différence de proportion de chaque allèle a été testée par une approche binomiale. Pour les deux études, en raison du nombre faible d'animaux permettant de valider la fréquence d'un SNP dans une population, nous avons filtré les ASE-SNPs avec les SNPs détectés dans la base de donnée dbSNP (Sherry *et al.*, 2001).

Dans la première étude, on a tenté d'identifier des variants causaux dans les régions régulatrices responsables du déséquilibre allélique. Dans ce but, tous les SNPs en LD avec un ASE-SNPs ont été recensés pour calculer la corrélation de Spearman du génotype de chacun de ces SNP avec le niveau d'expression du transcrit contenant le SNP respectif pour les 19 individus. Malheureusement, le nombre d'individus reste faible rendant le test peu puissant et tous les SNPs avec un génotype manquant pour au moins un individu ont été retirés pour éviter des biais. Pour pallier, le manque de puissance du test, il sera préférable d'augmenter le nombre d'individu testé à au moins 100. Or avec 100 individus, le risque d'avoir un génotype inconnu pour un des SNPs devient plus grand, il est donc nécessaire d'améliorer la couverture du génome séquencé en parallèle réduisant ainsi le nombre de SNPs éliminés. Enfin, une corrélation n'indique pas la causalité, ce qui signifie que la corrélation toute seule n'est pas suffisante pour prédire des variants causaux. Pour la deuxième étude, en raison du nombre vraiment faible d'individus, les corrélations n'ont pas été calculées. Une importante limite de l'approche ASE est la capacité de ne détecter des ASE que pour les SNP présents dans la population sous forme hétérozygote et avec une fréquence suffisante.

La détection d'expression de gènes spécifiques d'allèle à partir de données de RNA-Seq permet d'identifier des phénomènes d'empreinte parentale liés à un contrôle épigénétique (DeVeale *et al.*, 2012). Notre analyse ne permet pas d'identifier précisément les expressions allèle-spécifiques soumises à l'empreinte parentale ou non. Pouvoir identifier quel allèle provient de quel parent permettrait d'observer si l'allèle surexprimé provient toujours du même parent.

B.3. Tests et améliorations

B.3.1. Test du pipeline

Durant ma thèse, j'ai effectué une mission de 3 mois au Roslin Institute à Edimbourg dans l'équipe du professeur David Hume (voir B.3.3, page 119) financé par COST Action dans le cadre de FAANG (Functional Annotation of Animal Genetics). Cette mission m'a permis de tester mon pipeline d'analyse des ASE (Figure 23, page 50) sur d'autres échantillons et surtout sur d'autres espèces : le mouton et le buffle. Pour le buffle, les résultats ont été très peu concluants en raison de l'absence d'un assemblage de qualité à l'époque. Pour le mouton, il s'agissait d'une race croisée de père de race Texel et de mère de race Scottish Blackface. J'ai utilisé mon pipeline avec des données RNA-Seq de *biceps* de 3 brebis et 3 béliers ainsi que la séquence de leur génome entier. J'ai détecté près de 18 721 ASE-SNPs ce qui est assez impressionnant au regard des 3 369 ASE-SNPs détectés dans les échantillons musculaires des 6 vaches Holstein de la deuxième étude. Cette différence peut être due à une fréquence largement plus importante du phénomène ASE chez l'ovin, contrairement au bovin mais le plus probable découle de l'hybridation des moutons étudiés. En effet, étant d'une race croisée, la chance d'avoir des SNPs hétérozygotes est plus grande ce qui confirme que l'approche ASE est limitée par le nombre de SNPs hétérozygotes. Cette mission m'a permis aussi d'améliorer les performances du pipeline et d'améliorer les paramètres pour qu'il soit facilement utilisable sur différentes espèces. Malheureusement, aucun de ces ASE-SNPs détectés n'a été validé expérimentalement.

B.3.2. Améliorer la détection *in silico* des variants causaux candidats

Grâce aux deux études, on a montré que l'approche ASE pouvait faciliter l'identification de SNPs candidats régulateurs en *cis*. On sait toutefois qu'un ASE-SNP est rarement le variant causal. En parallèle, on a vu dans l'introduction (B.3.3, page 39) que les eGWAS (ou analyse eQTL) permettent la détection des effets en *cis* et en *trans* ainsi que la détection des *cis*-eQTL pouvait être complétée par les analyses ASE en caractérisant directement les variants agissant en *cis*. Les analyses ASE étant en effet plus puissantes grâce à un meilleur contrôle intra-individuel qui élimine ainsi les influences génétique en *trans* et les facteurs environnementaux (Pastinen, 2010).

Des études couplant les approches ASE et eQTLs ont déjà été réalisés, notamment chez la souris (Hasin-Brumshtein *et al.*, 2014) ou le porc (Maroilley *et al.*, 2017). Très récemment (Novembre 2018), l'analyse croisée des deux approches a été publiée chez le bovin par Khansefid *et al.* (2018). L'étude a été réalisée avec le transcriptome hépatique (37 échantillons) et musculaire (45 échan-

tillons) de taureaux Angus et avec le transcriptome hépatique et leucocytaire de 20 vaches Holstein. Dans cette étude, ils ont effectué des analyses eQTL et ASE mais aussi les ASE liés à une influence parentale, c'est à dire quand on détecte un déséquilibre allélique lié à une potentielle empreinte parentale (Figure 26). Ils différencient les tSNPs (transcript SNPs) qui présentent un déséquilibre allélique des dSNPs (driver SNPs) qui est le variant causant généralement dans la région régulatrice. Ils ont découvert que les dSNPs responsable de l'effet ASE était souvent des eQTLs locaux et donc probablement des *cis*-eQTLs. Ces dSNPs affectent souvent, mais pas toujours, l'expression des gènes dans de multiples tissus et (quand il l'affecte) l'allèle augmentant l'expression est généralement le même. Ils concluent que l'approche ASE détecte des phénomènes coïncidant avec des eQTL locaux, mais qu'il existe aussi des différences systématiques entre les SNPs détectés par les deux méthodes.

Dans l'objectif d'améliorer la détection des variants causaux à grande échelle, il est donc primordial de coupler les deux approches mais des améliorations peuvent encore être apportées. En étudiant, par exemple, les motifs de fixation des facteurs de transcription, on peut réduire la zone de recherche de variants régulant l'expression (Wang *et al.*, 2018). Il pourrait donc être intéressant de coupler les deux approches précédentes avec une détection de tous les SNPs dont la présence affecte des TFBS (Figure 27A). En parallèle, des SNPs dans la région 3'UTR peuvent altérer la fixation d'un microARN jouant aussi sur l'expression allèle-spécifique (Võsa *et al.*, 2015). Encore une fois, coupler les approches ASE et eQTL avec la détection des sites de fixation au microARN (Figure 27B) permettrait d'améliorer la détection des variants causaux à grande échelle.

B.3.3. D'autres méthodes de validation expérimentale

La méthode de pyroséquençage permet de quantifier une expression allèle-spécifique mais n'est pas très efficace pour une validation à grande échelle et pour valider les SNPs régulateurs. En revanche, l'approche MPRA permet d'analyser en parallèle plusieurs milliers de rSNPs candidats (voir B.4.1, page 42) et pour l'instant, elle n'a été uniquement utilisé pour tester des rSNPs humains (Tewhey *et al.*, 2016; Ulirsch *et al.*, 2016). Cela permettrait d'étudier un grand nombre de nos séquences allèle-spécifiques correspondant aux rSNPs qu'on souhaite tester.

Il existe une approche plus récente : le single-cell RNA sequencing (scRNA-seq Hedlund et Deng, 2018). L'objectif est d'étudier au niveau d'une seule cellule les *cis*-eQTLs ou les ASE (Van der Wijst *et al.*, 2018), afin d'étudier précisément ceux qui sont spécifique à un type de cellule, ce qui est fréquent (Brown *et al.*, 2013). Le principe repose sur le tri et l'isolement des cellules après dissociation d'un tissu. L'ARN est isolé pour être amplifié par PCR et l'ADNc ainsi amplifié est séquencé pour ensuite obtenir les profils d'expression pour chaque cellule individuellement (Figure 28).

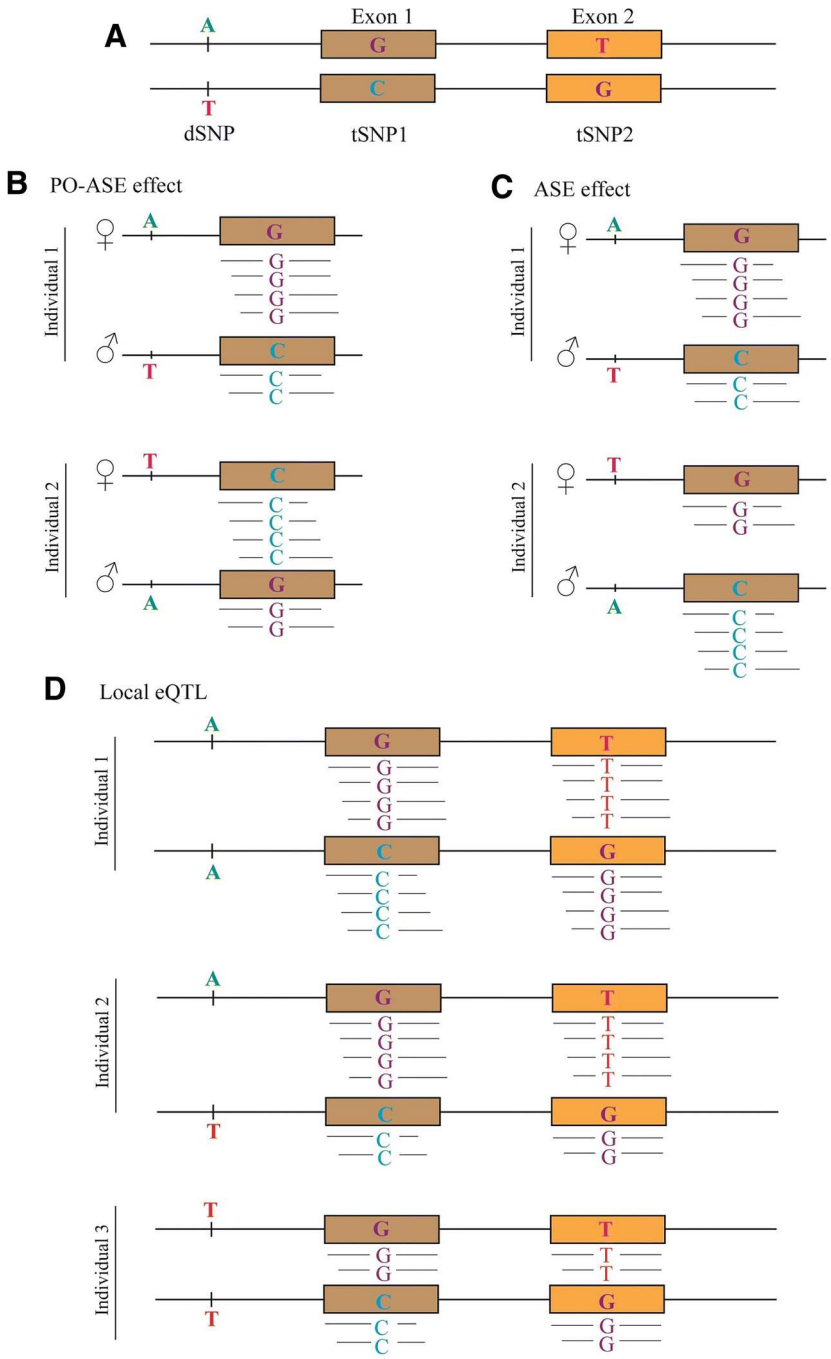

FIGURE 26 – Diagramme de l'expression allèle spécifique (ASE), de l'ASE lié à une influence parentale (PO-ASE) et du locus de caractères quantitatifs associé à l'expression (eQTL) d'après Khansefid *et al.* (2018). **A** Le gène a deux SNPs dans les exons ($tSNP_1$ et $tSNP_2$) dont l'expression peut être mesurée avec du RNA-Seq et un SNP dans la région promotrice local d'environ 50kb (dSNP). **B** Dans le cas des PO-ASE, l'allèle hérité de la mère (dans cet exemple) augmente l'expression de l'allèle du tSNP présent sur le chromosome maternel. **C** Quand le dSNP a un effet ASE, l'allèle A du dSNP (dans cet exemple) déclenche l'expression de l'allèle du tSNP sur le même chromosome. **D** Dans un alignement eQTL local, l'expression de l'allèle du dSNP est corrélé avec l'expression total de l'exon.

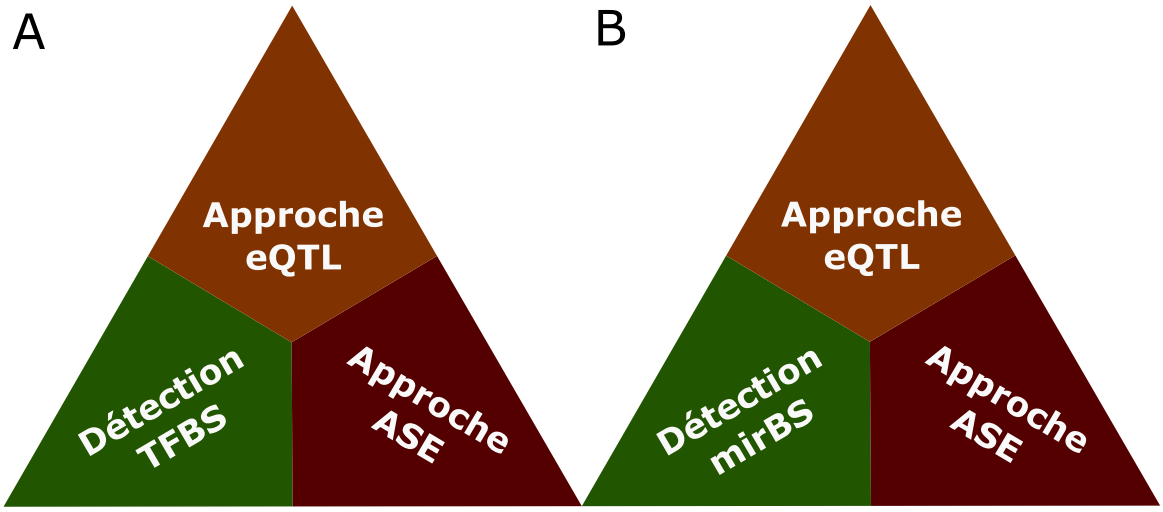

FIGURE 27 – Schéma de la détection des variants causaux en couplant les approches ASE et eQTL avec (A) la détection des TFBS (Transcription Factor Binding Site) et (B) la détection des mirBS (microRNA Binding Site).

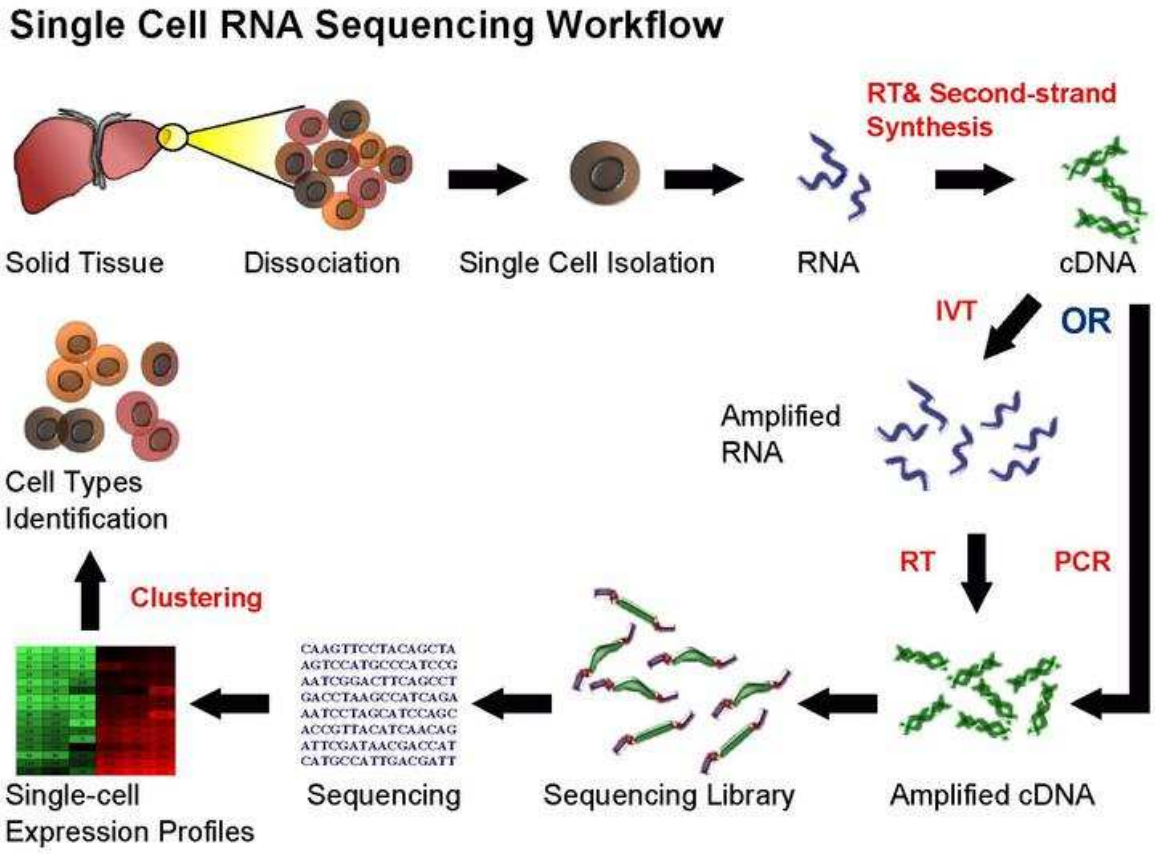

FIGURE 28 – Workflow d'analyse du single-cell RNA sequencing.

Bibliographie

Abascal, F., Juan, D., Jungreis, I., Martinez, L., Rigau, M., Rodriguez, J. M., Vazquez, J., et Tress, M. L. (2018). Loose ends : almost one in five human genes still have unresolved coding status. *Nucleic Acids Research*, 46(14) :7070–7084.

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M. J., Tuschl, T., et Margalit, H. (2005). Clustering and conservation patterns of human microRNAs. *Nucleic Acids Research*, 33 :2697–2706.

Amaral, P. P., Dinger, M. E., Mercer, T. R., et Mattick, J. S. (2008). The eukaryotic genome as an RNA machine. *Science*, 319(5871) :1787–1789.

Amengual, J., Guo, L., Strong, A., Madrigal-Matute, J., Wang, H. z., Kaushik, S., L Brodsky, J., J Rader, D., Maria Cuervo, A., et Fisher, E. (2018). Autophagy is required for sortilin-mediated degradation of Apolipoprotein B100. *Circulation Research*, 122 :568–582.

Amit, I., Garber, M., Chevrier, N., Leite, A. P., Donner, Y., Eisenhaure, T., et Regev, A. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating the differential response to pathogens. *Science*, 326 :257–263.

Anders, S., Pyl, P. T., et Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2) :166–169.

Anders, S., Reyes, A., et Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10) :2008–2017.

Annunziato, A. T. (2008). DNA packaging : Nucleosomes and chromatin. *Nature Education*, 1 :26.

Anttinen, H., Puistola, U., Pihlajaniemi, T., et Kivirikko, K. (1981). Differences between proline and lysine hydroxylations in their inhibitio by zinc or by ascorbate deficiency during collagen synthesis in various cell types. *Biochim. Biophys. Acta*, 674 :336–344.

Aphasizhev, R., Sbicego, S., Peris, M., Jang, S.-H., Aphasizheva, I., Simpson, A. M., Rivlin, A., et Simpson, L. (2002). Trypanosome mitochondrial 3'terminal uridylyl transferase (TUTase) : The key enzyme in U-insertion/deletion RNA editing. *Cell*, 108 :637 – 648.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., et Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123) :1074–1077.

Autran, D., Huanca-Mamani, W., et Vielle-Calzada, J.-P. (2005). Genomic imprinting in plants : the epigenetic version of an Oedipus complex. *Current Opinion in Plant Biology*, 8(1) :19 – 25.

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., et DePristo, M. A. (2013). From fastq data to high-confidence variant calls : The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43 :11.10.1–11.10.33.

Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P., et Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes & Development*, 22 :2773–2785.

Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, 26(12) :i318–i324.

Baralle, M. et Baralle, F. E. (2018). The splicing code. *Biosystems*, 164 :39 – 48.

Barlow, D. P. et Bartolomei, M. S. (2014). Genomic imprinting in mammals. *Cold Spring Harbor Perspectives in Biology*, 6(2).

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., et Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4) :823 – 837.

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., et Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4 :NOT deadenylase and DCP1 :DCP2 decapping complexes. *Genes & Development*, 20 :1885–1898.

Bell, A. et Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the IGF2 gene. *Nature*, 405 :482–5.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218) :53–59.

Berg, J., Tymoczko, J., et Stryer, L. (2002). *Many Enzymes Are Activated by Specific Proteolytic Cleavage*, chapter Section 10.5. New York : W H Freeman.

Bestor, T. H. (2000). The DNA methyltransferases of mammals. *Human Molecular Genetics*, 9(16) :2395–2402.

Bhan, A., Soleimani, M., et Mandal, S. S. (2017). Long noncoding RNA and cancer : A new paradigm. *Cancer Research*, 77(15) :3965–3981.

Bickhart, D. M., Hou, Y., Schroeder, S. G., Alkan, C., Cardone, M. F., Matukumalli, L. K., Song, J., Schnabel, R. D., Ventura, M., Taylor, J. F., Garcia, J. F., Van Tassell, C. P., Sonstegard, T. S., Eichler, E. E., et Liu, G. E. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research*, 22(4) :778–790.

Bird, A. P. (1986). CpG-rich islands and function of DNA methylation. *Nature*, 321 :209–13.

Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B. T., Krásný, L., Lee, J. K., Kaplan, C. D., Ebright, R. H., et Nickels, B. E. (2016). The mechanism of RNA 5' capping with NAD+, NADH, and desphospho-CoA. *Nature*, 535.

Blanc, V., Farré, J. C., Litvak, S., et Araya, A. (2002). Réécriture du matériel génétique : fonctions et mécanismes de l'édition de l'ARN. *médecine/sciences*, 18(2) :181–192.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., et Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, 326(5959) :1509–1512.

Bortfeldt, R., Schindler, S., Szafranski, K., Schuster, S., et Holste, D. (2008). Comparative analysis of sequence features involved in the recognition of tandem splice sites. *BMC Genomics*, 9.

Brockdorff, N., Ashworth, A., Kay, G. F., Cooper, P., Smith, S. M., McCabe, V. M., Norris, C., Penny, G. D., Patel, D. P., et Rastan, S. (1991). Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature*, 351 :329–331.

Broderick, J., Salomon, W., Ryder, S. P., Aronin, N., et Zamore, P. (2011). Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. *RNA*, 17 :1858–69.

Brown, C. D., Mangravite, L. M., et Engelhardt, B. E. (2013). Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLOS Genetics*, 9(8) :1–19.

Bártová, E., Krejčí, J., Harničarová, A., Galiová, G., et Kozubek, S. (2008). Histone modifications and nuclear architecture : A review. *Journal of Histochemistry & Cytochemistry*, 56 :711–721.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, 25(18) :1915–1927.

Cantara, W. A., Crain, P. F., Rozenski, J., McCloskey, J. A., Harris, K. A., Zhang, X., Vendeix, F. A. P., Fabris, D., et Agris, P. F. (2011). The RNA modification database, RNAMDB : 2011 update. *Nucleic Acids Research*, 39 :D195–D201.

Carrel, L. et Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, 434 :400–404.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., Forrest, A., Carninci, P., Biffo, S., Stupka, E., et Gustincich, S. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, 491 :454–457.

Casci, T. (2010). SNPs that come in threes. *Nature Reviews Genetics*, 11 :8.

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., et Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, 16(1) :195.

Cell Signaling Technology (2018). Histone modification table description.

Chamberlain, A. J., Vander Jagt, C., Hayes, B., Khansefid, M., Marett, L., Millen, C., et Goddard, M. E. (2015). Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics*, 16 :993.

Chen, H., Kronengold, J., Yan, Y., Gazula, V.-R., R Brown, M., Ma, L., Ferreira, G., Yang, Y., Bhattacharjee, A., Sigworth, F., Salkoff, L., et K Kaczmarek, L. (2009). The N-terminal domain of slack determines the formation and trafficking of slick/slack heteromeric sodium-activated potassium channels. *The Journal of neuroscience*, 29 :5654–65.

Chen, J., Rozowsky, J., Galeev, T. R., Harmanci, A., Kitchen, R., Bedford, J., et Gerstein, M. (2016). A uniform survey of allele-specific binding and expression over 1000-genomes-project individuals. *Nature Communications*, 7 :11101.

Chen, S., Habib, G., Yang, C., Gu, Z., Lee, B., Weng, S., Silberman, S., Cai, S., Deslypere, J., et Rosseneu, M. (1987). Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science*, 238 :363–366.

Cheng, H., Dufu, K., Lee, C.-S., Hsu, J. L., Dias, A., et Reed, R. (2006). Human mRNA Export Machinery Recruited to the 5' End of mRNA. *Cell*, 127 :1389–1400.

Chial, H. (2008). DNA sequencing technologies key to the Human Genome Project. *Nature Education*, 1 :219.

Chitwood, J. L., Rincon, G., Kaiser, G. G., Medrano, J. F., et Ross, P. J. (2013). RNA-Seq analysis of single bovine blastocysts. *BMC Genomics*, 14 :350.

Choi, J.-W., Choi, B.-H., Lee, S.-H., Lee, S.-S., Kim, H.-C., Yu, D., et Lim, D. (2015). Whole-genome resequencing analysis of hanwoo and yanbian cattle to identify genome-wide snps and signatures of selection. *Molecules and Cells*, 38(5) :466–473.

Choi, K. M., Barash, I., et Rhoads, R. E. (2004). Insulin and prolactin synergistically stimulate beta-casein messenger ribonucleic acid translation by cytoplasmic polyadenylation. *Molecular Endocrinology*, 18 :1670–1686.

Cibulskis, K., Sohn, L. M., Carter, S. L., Sivachenko, A. Y., Jaffe, D. B., Sougnez, C. L., Gabriel, S. B., Meyerson, M. L., Lander, E. S., et Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3) :213–9.

Comb, M. et Goodman, H. M. (1990). CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. *Nucleic acids research*, 18 :3975–82.

Cooper, D. W., Vandeberg, J. L., Sharman, G. B., et Poole, W. E. (1971). Phosphoglycerate kinase polymorphism in kangaroos provides further evidence for paternal X inactivation. *Nature New Biology*, 230(13) :155–157.

Courtney, D. G., Moore, J. E., Atkinson, S. D., Maurizi, E., Allen, E. H. A., Pedrioli, D. M. L., McLean, W. H. I., Nesbit, M. A., et Moore, C. B. T. (2016). CRISPR/Cas9 DNA cleavage at SNP-derived PAM enables both in vitro and in vivo KRT12 mutation-specific targeting. *Gene Therapy*, 23 :108–112.

Crowley, J. J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I. K., Kim, Y., Wang, J. R., Morgan, A. P., Calaway, J. D., Aylor, D. L., Yun, Z., Bell, T. A. H., et de Villena, F. P.-M. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics*, 47 :353–360.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., et Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24) :3207–3212.

Delaney, C., Garg, S. K., et Yung, R. (2015). *Analysis of DNA Methylation by Pyrosequencing*, pages 249–264. Springer New York.

DePristo, M. A., Banks, E. D., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Angel, G. D., Rivas, M. A., Hanna, M. C., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. M., et Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5) :491–8.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J., et Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs : Analysis of their gene structure, evolution, and expression. *Genome Research*, 22(9) :1775–1789.

DeVeale, B., van der Kooy, D., et Babak, T. (2012). critical evaluation of imprinted gene expression by RNA–Seq : a new perspective. *PLOS Genetics*, 8(3) :1–12.

Di Liegro, C., Schiera, G., et Di Liegro, I. (2014). Regulation of mRNA transport, localization and translation in the nervous system of mammals (Review). *International Journal of Molecular Medicine*, 33 :747–762.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., et Gingeras, T. R. (2013). STAR : ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1) :15–21.

Dominissini, D. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, 485 :201–6.

Dong, A., Yoder, J. A., Zhang, X., Zhou, L., Bestor, T. H., et Cheng, X. (2001). Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. *Nucleic Acids Research*, 29(2) :439–448.

Dou, Y., Fox-Walsh, K. L., Baldi, P. F., et Hertel, K. J. (2006). Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA*, 12 :2047–2056.

Drapkin, R., Reardon, J., Ansari, A., Huang, J.-C., Zawel, L., Ahn, K., Sancar, A., et Reinberg, D. (1994). Dual role of TFIIH in DNA excision repair and in transcription by RNA polymerase II. *Nature*, 368 :769–772.

Dupont, C., Armant, D., et Brenner, C. (2009). Epigenetics : Definition, mechanisms and clinical perspective. *Seminars in reproductive medicine*, 27 :351–7.

Eipper, B. A., Milgram, S. L., Jean Husten, E., Yun, H.-Y., et Mains, R. E. (1993). Peptidylglycine alpha-amidating monooxygenase : A multifunctional protein with catalytic, processing, and routing domains. *Protein Science*, 2 :489–497.

El-Osta, A. et Wolffe, A. (2001). DNA methylation and histone deacetylation in the control of gene expression : Basic biochemistry to human development and disease. *Gene Expression*, 9 :63–75.

Esteve-Codina, A., Kofler, R., Palmieri, N., Bussotti, G., Notredame, C., et Pérez-Enciso, M. (2011). Exploring the gonad transcriptome of two extreme male pigs with RNA-Seq. *BMC Genomics*, 12 :552.

Esvelt, K. M. et Wang, H. H. (2013). Genome-scale engineering for systems and synthetic biology. *Molecular Systems Biology*, 9(1).

Everts-van der Wind, A., Kata, S. R., Band, M. R., Rebeiz, M., Larkin, D. M., Everts, R. E., Green, C. A., Liu, L., Natarajan, S., Goldammer, T., Lee, J. H., McKay, S., Womack, J. E., et Lewin, H. A. (2004). A 1463 gene cattle–human comparative map with anchor points defined by human genome sequence coordinates. *Genome Research*, 14(7) :1424–1437.

Fear, J. M., León-Novelo, L. G., Morse, A. M., Gerken, A. R., Van Lehmann, K., Tower, J., et McIntyre, L. M. (2016). Buffering of genetic regulatory networks in drosophila melanogaster. *Genetics*, 203 :1177–1190.

Fossat, N. et Tam, P. P. L. (2014). Re-editing the paradigm of Cytidine (C) to Uridine (U) RNA editing. *RNA Biology*, 11(10) :1233–1237.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., et Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19 :92–105.

Fritz, S. et Boris-Lawrie, K. (2015). The RNPs of eukaryotic translation control. *Trends in Cell and Molecular Biology*, 10 :105–130.

Garcia-Lopez, J., Brieño-Enríquez, M., et del Mazo, J. (2013). MicroRNA biogenesis and variability. *Biomolecular concepts*, 4 :367–380.

Geiger, J. H., Hahn, S., Lee, S., et Sigler, P. B. (1996). Crystal structure of the yeast TFIIA/TBP/DNA complex. *Science*, 272(5263) :830–836.

Geisler, S. et Coller, J. (2013). RNA in unexpected places : Long non-coding RNA functions in diverse cellular contexts. *Nature reviews. Molecular cell biology*, 14 :699–712.

Ghazanfar, S., Vuocolo, T., Morrison, J. L., Nicholas, L. M., McMillen, I. C., Yang, J. Y. H., et Tellam, R. L. (2017). Gene expression allelic imbalance in ovine brown adipose tissue impacts energy homeostasis. *PlOS ONE*, 12 :e0180378.

Gong, F. et Miller, K. M. (2013). Mammalian dna repair : Hats and hdacs make their mark through histone acetylation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 750(1) :23 – 30.

Gonzalez, K. D., Hill, K. A., Li, K., Li, W., Scaringe, W. A., Wang, J.-C., Gu, D., et Sommer, S. S. (2007). Somatic microindels : analysis in mouse soma and comparison with the human germline. *Human Mutation*, 28(1) :69–80.

Gott, J. M. et Emeson, R. B. (2000). Functions and mechanisms of RNA editing. *Annual Review of Genetics*, 34(1) :499–531.

Graveley, B. (2001). Alternative splicing : Increasing diversity in the proteomic world. *Trends in genetics*, 17 :100–7.

Greer, E. L. et Shi, Y. (2012). Histone methylation : a dynamic mark in health, disease and inheritance. *Nature Reviews Genetics*, 13 :343–357.

Grigoryev, D. N., Cheranova, D. I., Chaudhary, S., Heruth, D. P., Zhang, L. Q., et Ye, S. Q. (2015). Identification of new biomarkers for acute respiratory distress syndrome by expression-based genome-wide association study. *BMC Pulmonary Medicine*, 15 :95.

Grummt, I. (1998). Regulation of mammalian ribosomal gene transcription by RNA polymerase I. *Progress in Nucleic Acid Research and Molecular Biology*, 62 :109–154.

Gu, F. et Wang, X. (2015). Analysis of allele specific expression – a survey. *Tsinghua Science and Technology*, 20(5) :513–529.

Gu, S. et Kay, M. A. (2010). How do miRNAs mediate translational repression ? *Silence*, 1 :11.

Guo, Y. et Jamison, D. C. (2005). The distribution of SNPs in human gene regulatory regions. *BMC Genomics*, 6(1) :140.

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M.-C., Hung, T., et Chang, H. Y. (2010). Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464 :1071–1076.

Ha, M. et Narry Kim, V. (2014). Regulation of microRNA biogenesis. *Nature reviews. Molecular cell biology*, 15 :509–524.

Haley, C. et De Koning, D. J. (2006). Genetical genomics in livestock : potentials and pitfalls. *Animal Genetics*, 37(s1) :10–12.

Han, Y., Yang, Y.-N., Yuan, H.-H., Zhang, T.-T., Sui, H., Wei, X.-L., Liu, L., Huang, P., Zhang, W.-J., et Bai, Y.-X. (2014). UCA1, a long non-coding RNA up-regulated in colorectal cancer influences cell proliferation, apoptosis and cell cycle distribution. *Pathology*, 46 :396–401.

Hang, L. (2012). *Analyse de l'efficacité de la régulation par les microARN.* PhD thesis, Université Paris Sud - Paris XI.

Hansen, K. D., Irizarry, R. A., et WU, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2) :204–216.

Hasin-Brumshtein, Y., Hormozdiari, F., Martin, L., van Nas, A., Eskin, E., Lusis, A. J., et Drake, T. A. (2014). Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics*, 15(1) :471.

Hayden, E. C. (2014). Technology : The $1,000 genome. *Nature*, 507 :294–295.

Hedlund, E. et Deng, Q. (2018). Single-cell RNA sequencing : Technical advancements and biological applications. *Molecular Aspects of Medicine*, 59 :36 – 46.

Herr, A. J., Jensen, M. B., Dalmay, T., et Baulcombe, D. C. (2005). RNA Polymerase IV Directs Silencing of Endogenous DNA. *Science*, 308 :118–120.

Hershko, A. et Ciechanover, A. (1998). The ubiquitin system. *Annual Review of Biochemistry*, 67 :425–479.

Higgins, M., Fitzsimons, C., C. McClure, M., McKenna, C., Conroy, S., Kenny, D., Mcgee, M., Waters, S., et W. Morris, D. (2018). GWAS and eQTL analysis identifies a SNP associated with both residual feed intake and GFRA2 expression in beef cattle. *Scientific Reports*, 8.

Hsu, P. D., Lander, E. S., et Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell*, 157(6) :1262–1278.

Hsu, T. C. et Arrighi, F. E. (1971). Distribution of constitutive heterochromatin in mammalian chromosomes. *Chromosoma*, 34(3) :243–253.

Hunt, R. W., Mathelier, A., del Peso, L., et Wasserman, W. W. (2014). Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics*, 15(1) :472.

Ionita-Laza, I., Rogers, A. J., Lange, C., Raby, B. A., et Lee, C. (2009). Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*, 93(1) :22 – 26.

Irvine, R. A., Lin, I. G., et Hsieh, C.-L. (2002). DNA methylation has a local effect on transcription and histone acetylation. *Molecular and Cellular Biology*, 22(19) :6689–6696.

Jackson, R., U.T. Hellen, C., et V Pestova, T. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature reviews. Molecular cell biology*, 11 :113–27.

Jenuwein, T. et Allis, C. D. (2001). Translating the histone code. *Science*, 293(5532) :1074–80.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., et Charpentier, E. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096) :816–821.

Jones, P. A. (2012). Functions of DNA methylation : Islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13 :484–92.

Ju, Z., Wang, C., Wang, X., Yang, C., Sun, Y., Jiang, Q., Wang, F., Li, M., Zhong, J., et Huang, J. (2015). Role of an SNP in alternative splicing of bovine NCF4 and mastitis susceptibility. *PLOS ONE*, 10(11) :1–15.

Jugder, B.-E., Welch, J., Braidy, N., et Marquis, C. P. (2016). Construction and use of a *Cupriavidus necator* H16 soluble hydrogenase promoter (P_{SH}) fusion to *gfp* (green fluorescent protein). *PeerJ*, 4 :e2269.

Jung, H. J. et Suh, Y. (2015). Regulation of IGF -1 signaling by microRNAs. *Frontiers in Genetics*, 5 :472.

Jura, N., Scotto-Lavino, E., Sobczyk, A., et Bar-Sagi, D. (2006). Differential modification of ras proteins by ubiquitination. *Molecular Cell*, 21 :679 – 687.

Jurica, M. et Moore, M. (2003). Pre-mRNA splicing : awash in a sea of proteins. *Molecular cell*, 12 :5–14.

Kabanova, S., Kleinbongard, P., Volkmer, J., Andrée, B., Kelm, M., et Jax, T. (2009). Gene expression analysis of human red blood cells. *International journal of medical sciences*, 6 :156–9.

Kapp, L. D. et Lorsch, J. R. (2004). The molecular mechanics of eukaryotic translation. *Annual Review of Biochemistry*, 73 :657–704.

Kendziorski, C. et Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome*, 17(6) :509–517.

Keren, H., Lev-Maor, G., et Ast, G. (2010). Alternative splicing and evolution : diversification, exon definition and function. *Nature Reviews Genetics*, 11 :345–55.

Khansefid, M., Pryce, J. E., Bolormaa, S., Chen, Y., Millen, C. A., Chamberlain, A. J., Vander Jagt, C. J., et Goddard, M. E. (2018). Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. *BMC Genomics*, 19(1) :793.

Khoury, G., C. Baliban, R., et A. Floudas, C. (2011). Proteome-wide post-translational modification statistics : frequencyanalysis and curation of the swiss-prot database. *Scientific Reports*, 1.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., et Salzberg, S. L. (2013). TopHat2 : accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4) :R36.

Kim, Y. G., Cha, J., et Chandrasegaran, S. (1996). Hybrid restriction enzymes : zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences*, 93(3) :1156–1160.

Knight, J. C. (2014). Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Medicine*, 6(10) :92.

Kolhekar, A. S., Roberts, M. S., Jiang, N., Johnson, R. C., Mains, R. E., Eipper, B. A., et Taghert, P. H. (1997). Neuropeptide amidation in drosophila : Separate genes encode the two enzymes catalyzing amidation. *Journal of Neuroscience*, 17 :1363–1376.

Koo, J., Kim, Y., Kim, J., Yeom, M., Lee, I. C., et Nam, H. G. (2007). A GUS/luciferase fusion reporter for plant gene trapping and for assay of promoter activity with luciferin-dependent control of the reporter protein stability. *Plant and Cell Physiology*, 48(8) :1121–1131.

Kreutz, M., Hochstein, N., Kaiser, J., Narz, F., et Peist, R. (2013). Pyrosequencing : Powerful and quantitative sequencing technology. *Current Protocols in Molecular Biology*, 104(1) :7.15.1–7.15.23.

Krishnamurthy, S. et Hampsey, M. (2009). Eukaryotic transcription initiation. *Current Biology*, 19 :R153–R156.

Krishnan, V. et Zeichner, S. L. (2004). Host cell gene expression during human immunodeficiency virus type 1 latency and reactivation and effects of targeting genes that are differentially expressed in viral latency. *Journal of Virology*, 78(17) :9458–9473.

Krol, J., Loedige, I., et Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics*, 11 :597–610.

Kulis, M., Queiros, A. C., Beekman, R., et Martin-Subero, J. I. (2013). Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(11) :1161–1174.

Lagarrigue, S., Martin, L., Hormozdiari, F., Roux, P.-F., Pan, C. anf van Nas, A., et Lusis, A. J. (2013). Analysis of allele-specific expression in mouse liver by RNA-Seq : A comparison with cis-eQTL identified using genetic linkage. *Genetics*, 195 :1157–1166.

Langmead, B. et Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9 :357–359.

Lee, C.-Y., Chiu, Y.-C., Wang, L.-B., Kuo, Y.-L., Chuang, E. Y., Lai, L.-C., et Tsai, M.-H. (2013). Common applications of next-generation sequencing technologies in genomic research. *Translational Cancer Research*, 2(1).

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W. R., et Schatz, M. (2016). Third-generation sequencing and the future of genomics. *bioRxiv*.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., et Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*, 23 :4051–4060.

Letaief, R., Rebours, E., Grohs, C., Meersseman, C., Fritz, S., Trouilh, L., Esquerré, D., Barbieri, J., Klopp, C., Philippe, R., Blanquet, V., Boichard, D., Bousaha, M., et Rocha, D. (2017). Identification of copy number variation in french dairy and beef breeds using next-generation sequencing. *Genetics Selection Evolution*, 49(1) :77.

Li, E., Beard, C., et Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, 366 :362–5.

Li, G., Bahn, J. H., Lee, J.-H., Peng, G., Chen, Z., Nelson, S. F., et Xiao, X. (2012). Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research*, 40(13) :e104.

Li, H. et Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14) :1754–1760.

Li, J. B., Levanon, E. Y., Yoon, J.-K., Aach, J., Xie, B., LeProust, E., Zhang, K., Gao, Y., et Church, G. M. (2009). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science*, 324 :1210–1213.

Li, S. et Mason, C. E. (2014). The pivotal regulatory landscape of RNA modifications. *Annual Review of Genomics and Human Genetics*, 15(1) :127–150.

Liang, F., Han, M., Romanienko, P. J., et Jasin, M. (1998). Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *PNAS*, 95(9) :5172–5177.

Locker, N., Chamond, N., et Sargueil, B. (2011). A conserved structure within the HIV gag open reading frame that controls translation initiation directly recruits the 40S subunit and eIF3. *Nucleic Acids Research*, 39(6) :2367–2377.

Lonsdale, J. T., Thomas, J. A., Salvatore, M., Phillips, R. A., Lo, E., Shad, S., Hasz, R. D., Walters, G. D., García, F., Young, N. S., Foster, B. A., Moser, M., Karasik, E., Gillard, B. M., Ramsey, K. D., Sullivan, S. B., et Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45 :580–585.

Lopdell, T. J., Tiplady, K., Struchalin, M., Johnson, T. J. J., Keehan, M., Sherlock, R., et Littlejohn, M. D. (2017). DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. *BMC Genomics*, 18 :968.

Love, M. I., Huber, W., et Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12) :550.

Lytle, J. R., Wu, L., et Robertson, H. D. (2002). Domains on the hepatitis C virus internal ribosome entry site for 40S subunit binding. *RNA*, 8(8) :1045–1055.

Ma, L., Bajic, V. B., et Zhang, Z. (2013). On the classification of long non-coding RNAs. *RNA Biology*, 10(6) :924–933.

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., et Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070) :823–828.

Machnicka, M. A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K. M., Helm, M., Bujnicki, J. M., et Grosjean, H. (2013). MODOMICS : a database of RNA modification pathways—2013 update. *Nucleic Acids Research*, 41(D1) :D262–D267.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., et Bemben, L. A.and Rothberg, J. M. (2005). Genome sequencing in open microfabricated high density picoliter reactors. *Nature*, 437(7057) :376–380.

Maroilley, T., Lemonnier, G., Lecardonnel, J., Esquerré, D., Ramayo-Caldas, Y., Mercat, M. J., et Estellé, J. (2017). Deciphering the genetic regulation of peripheral blood transcriptome in pigs through expression genome-wide association study and allele-specific expression analysis. *BMC Genomics*, 18.

Maunakea, A., Chepelev, I., Cui, K., et Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell research*, 23 :1256–1269.

McCullough, A. J. et Berget, S. M. (1997). G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Molecular and Cellular Biology*, 17 :4562–4571.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et DePristo, M. A. (2010). The Genome Analysis Toolkit : A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. *Genome Research*, 20 :1297–1303.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M., et Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, 19(9) :1527–1541.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., et Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1) :122.

McLoughlin, K. E., Nalpas, N. C., Rue-Albrecht, K., Browne, J. A., Magee, D. A., Killick, K. E., Park, S. D. E., Hokamp, K., Meade, K. G., O'Farrelly, C., Gormley, E., Gordon, S. V., et MacHugh, D. E. (2014). Rna-seq transcriptional profiling of peripheral blood leukocytes from cattle infected with mycobacterium bovis. *Frontiers in Immunology*, 5 :396.

McManus, C. J. et Graveley, B. R. (2008). Getting the message out. *Molecular cell*, 31(1) :4–6.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., et Tuschl, T. (2004). Human Argonaute2 Mediates RNA Cleavage Targeted by miRNAs and siRNAs. *Molecular Cell*, 15 :185–197.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., G Callan, C., B Kinney, J., Kellis, M., S Lander, E., et S Mikkelsen, T. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology*, 30 :271–7.

Melnikov, A., Zhang, X., Rogov, P., Wang, L., et S Mikkelsen, T. (2014). Massively parallel reporter assays in cultured mammalian cells. *Journal of visualized experiments*, 90.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., et Jaffrey, S. R. (2012). Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons. *Cell*, 149(7) :1635 – 1646.

Miao, Y.-R., Liu, W., Zhang, Q., et Guo, A.-Y. (2018). lncRNASNP2 : an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Research*, 46(D1) :D276–D280.

Miglani, G. (2014). *Gene Expression*. Alpha Science.

Mills, J. D., Kawahara, Y., et Janitz, M. (2013). Strand-specific RNA-Seq provides greater resolution of transcriptome profiling. *Current Genomics*, 14(3) :173–181.

Mosesson, M. W. (2005). Fibrinogen and fibrin structure and functions. *Journal of Thrombosis and Haemostasis*, 3 :1894–1904.

Mullaney, J. M., Mills, R. E., Pittard, W. S., et Devine, S. E. (2010). Small insertions and deletions (indels) in human genomes. *Human Molecular Genetics*, 19(R2) :R131–R136.

Mundade, R., Ozer, H. G., Wei, H., Prabhu, L., et Lu, T. (2014). Role of ChIP-Seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, 13(18) :2847–2852.

Muráni, E., Ponsuksili, S., Srikanchai, T., Maak, S., et Wimmers, K. (2009). Expression of the porcine adrenergic receptor beta 2 gene in longissimus dorsi muscle is affected by cis-regulatory DNA variation. *Animal Genetics*, 40 :80–89.

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. *Annual Review of Biochemistry*, 79(1) :321–349.

Okano, M., Bell, D. W., Haber, D. A., et Li, E. (1999). DNA methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3) :247 – 257.

Okonechnikov, K., Conesa, A., et García-Alcalde, F. (2016). Qualimap 2 : advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32 :292–294.

Orphanides, G., Lagrange, T., et Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes & Development*, 10(21) :2657–2683.

Osorio, J. S. et Bionaz, M. (2017). Plasmid transfection in bovine cells : Optimization using a realtime monitoring of green fluorescent protein and effect on gene reporter assay. *Gene*, 626 :200 – 208.

Pai, A. A., Pritchard, J. K., et Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLOS Genetics*, 11 :1–8.

Park, P. J. (2009). ChIP-Seq : advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10 :669–680.

Pastinen, T. (2010). Genome-wide allele-specific analysis : insights into regulatory variation. *Nature Reviews Genetics*, 11 :533–538.

Patient, R. K. (1990). Control of gene expression : Tissue-specific expression. *Current Opinion in Biotechnology*, 1(2) :151 – 158.

Pearson, G., Robinson, F., Gibson, T. B., Xu, B.-e., Karandikar, M., Berman, K., et Cobb, M. H. (2001). Mitogen-activated protein (MAP) kinase pathways : Regulation and physiological functions. *Endocrine Reviews*, 22(2) :153–183.

Pennisi, E. (2010). Semiconductors inspire new sequencing technologies. *Science*, 327(5970) :1190–1190.

Pillai, R. S., Bhattacharyya, S. N., et Filipowicz, W. (2007). Repression of protein synthesis by miRNAs : how many mechanisms ? *Trends in cell biology*, 17 3 :118–26.

Pinto, I., Ware, D. E., et Hampsey, M. (1992). The yeast SUA7 gene encodes a homolog of human transcription factor TFIIB and is required for normal start site selection in vivo. *Cell*, 68(5) :977 – 988.

Pohl, M., Bortfeldt, R. H., Grützmann, K., et Schuster, S. (2013). Alternative splicing of mutually exclusive exons - a review. *Bio Systems*, 114 :31–8.

Ponsuksili, S., Murani, E., Trakooljul, N., Schwerin, M., et Wimmers, K. (2014). Discovery of candidate genes for muscle traits based on GWAS supported by eQTL-analysis. *International journal of biological sciences*, 10(3) :327–37.

Ponting, C. P., Oliver, P. L., et Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell*, 136 :629–41.

Pyle, A. M. (2016). Group II intron self-splicing. *Annual Review of Biophysics*, 45(1) :183–205.

Rahman, M. et Sadygov, R. G. (2017). Predicting the protein half-life in tissue from its cellular properties. *PLOS ONE*, 12 :1–15.

Ramachandran, H., Herfurth, K., Grosschedl, R., Schäfer, T., et Walz, G. (2015). SUMOylation blocks the ubiquitin-mediated degradation of the nephronophthisis gene product Glis2/NPHP7. *PLOS ONE*, 10 :1–16.

Ramaswami, G. et Li, J. B. (2014). RADAR : a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research*, 42 :D109–D113.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447 :425–432.

Reik, W. et Walter, J. (2001). Genomic imprinting : parental influence on the genome. *Nature Reviews Genetics*, 2 :21–32.

Renkawitz, J., Lademann, C., et Jentsch, S. (2014). Mechanisms and principles of homology search during recombination. *Nature reviews. Molecular cell biology.*, 15 :369–383.

Ribeiro-dos Santos, A. M., da Silva, V. L., de Souza, J. E., et de Souza, S. J. (2015). Populational landscape of indels affecting transcription factor-binding sites in humans. *BMC Genomics*, 16(1) :536.

Rice, J. C. et Allis, C. D. (2001). Histone methylation versus histone acetylation : new insights into epigenetic regulation. *Current Opinion in Cell Biology*, 13(3) :263 – 273.

Robert, M.-F., Morin, S., Beaulieu, N., Gauthier, F., Chute, I. C., Barsalou, A., et Macleod, A. R. (2003). DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nature Genetics*, 33 :61–65.

Roberts, S. G., Ha, I., Maldonado, E., Reinberg, D. F., et Green, M. R. (1993). Interaction between an acidic activator and transcription factor TFIIB is required for transcriptional activation. *Nature*, 363 :741–744.

Rockman, M. V. et Kruglyak, L. (2006). Genetics of global gene expression. *Nature reviews. Genetics*, 7 :862–872.

Ronaghi, M., Uhlén, M., et Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281(5375) :363–365.

Rosonina, E., Kaneko, S., et Manley, J. L. (2006). Terminating the transcript : breaking up is hard to do. *Genes & Development*, 20 :1050–1056.

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., Hirst, M., Turashvili, G., Oloumi, A., Marra, M. A., Aparicio, S., et Shah, S. P. (2012). JointSNVMix : a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, 28(7) :907–913.

Royo, J. L. et Galán, J. J. (2009). Pyrosequencing for SNP genotyping. *Methods in Molecular Biology*, 578 :123–33.

Royo, J. L., Hidalgo, M., et Ruiz, A. (2007). Pyrosequencing protocol using a universal biotinylated primer for mutation detection and SNP genotyping. *Nature protocols*, 2 :1734–9.

Ruby, J. G., Jan, C. H., et Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448 :83–86.

Sabbagh, U., Mullegama, S., et Wyckoff, G. J. (2016). Identification and evolutionary analysis of potential candidate genes in a human eating disorder. *BioMed Research International*, 2016 :1–11.

Sakabe, N. J. et de Souza, S. J. (2007). Sequence features responsible for intron retention in human. *BMC Genomics*, 8.

Saletore, Y., Chen-Kiang, S., et Mason, C. E. (2013). Novel RNA regulatory mechanisms revealed in the epitranscriptome. *RNA Biology*, 10(3) :342–346.

Sammeth, M., Foissac, S., et Guigó, R. (2008). A general definition and nomenclature for alternative splicing events. *PLOS Computational Biology*, 4 :1–14.

Sandmann, S., de Graaf, A. O., Karimi, M., van der Reijden, B. A., EvaHellström, Lindberg, Jansen, J., et Dugas, M. (2017). Evaluating variant calling tools for non-matched next-generation sequencing data. *Scientific Reports*, 7.

Sanger, F. et Coulson, A. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3) :441 – 448.

Sasaki, H. et Matsui, Y. (2008). Epigenetic events in mammalian germ-cell development : Reprogramming and beyond. *Nature reviews Genetics*, 9 :129–40.

Satya, R. V., Zavaljevski, N., et Reifman, J. (2012). A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Research*, 40(16) :e127.

Saunders, M. A., Liang, H., et Li, W.-H. (2007). Human polymorphism at microRNAs and microRNA target sites. *Proceedings of the National Academy of Sciences*, 104(9) :3300–3305.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, Peter S. Mao, M., Stoughton, R. B., et Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 297–302 :422.

Schadt, E. E., Turner, S., et Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2) :R227–R240.

Schwarz, F. et Aebi, M. (2011). Mechanisms and principles of n-linked protein glycosylation. *Current Opinion in Structural Biology*, 21 :576 – 582.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., et Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683) :525–528.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et Friedman, N. (2003). Module networks : identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34 :166–176.

Serre, D., Gurd, S., Ge, B., Sladek, R., Sinnett, D., Harmsen, E., Bibikova, M., Chudin, E., Barker, D. L., Dickinson, T., Fan, J.-B., et Hudson, T. J. (2008). Differential allelic expression in the human genome : A robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLOS Genetics*, 4(2) :1–16.

Shen, L., Song, C.-X., He, C., et Zhang, Y. (2014). Mechanism and function of oxidative reversal of DNA and RNA methylation. *Annual Review of Biochemistry*, 83(1) :585–614.

Sherry, S. T., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et Sirotkin, K. (2001). dbSNP : the NCBI database of genetic variation. *Nucleic acids research*, 29 :308–11.

Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Duchateau, P., et Pâques, F. (2011). Meganucleases and other tools for targeted genome engineering : Perspectives and challenges for gene therapy. *Current Gene Therapy*, 11(1) :11–27.

Skelly, D. A., Ronald, J., et Akey, J. M. (2009). Inherited variation in gene expression. *Annual Review of Genomics and Human Genetics*, 10(1) :313–332.

Smith, C. et Nadal-Ginard, B. (1989). Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location : Implications for constitutive splicing. *Cell*, 56 :749–58.

Smith, R. M., Webb, A., Papp, A. C., Newman, L. C., Handelman, S. K., Suhy, A., Mascarenhas, R., Oberdick, J., et Sadee, W. (2013). Whole transcriptome RNA-Seq allelic expression in human brain. *BMC Genomics*, 14(1) :571.

Soldner, F., Stelzer, Y., Shivalila, C. S., Abraham, B. J., Latourelle, J. C., Barrasa, M. I., Goldmann, J., Myers, R. H., Young, R. A., et Jaenisch, R. (2016). Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature*, 533 :95–99.

Spencer, C. A. et Groudine, M. (1990). Transcription elongation and eukaryotic gene regulation. *Oncogene*, 5(6) :777—785.

Springer, N. M. et Stupar, R. M. (2007). Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *The Plant Cell*, 19(8) :2391–2402.

Stasevich, T. J., Hayashi-Takanaka, Y., Sato, Y., Maehara, K., Ohkawa, Y., Sakata-Sogawa, K., Tokunaga, M., Nagase, T., Nozaki, N., McNally, J. G., et Kimura, H. (2014). Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature*, 516 :272–275.

Taft, R. J., Glazov, E. A., Lassmann, T., Hayashizaki, Y., Carninci, P., et Mattick, J. S. (2009). Small RNAs derived from snoRNAs. *RNA*, 15(7) :1233–1240.

Teixeira, A., Tahiri-Alaoui, A., West, S., Thomas, B., Ramadass, A., Martianov, I., Dye, M., James, W., J Proudfoot, N., et Akoulitchev, A. (2004). Autocatalytic RNA cleavage in the human beta-globin pre-mRNA promotes transcription termination. *Nature*, 432 :526–30.

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Young, S., Reilly, S., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S., et Sabeti, P. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, 165 :1519–1529.

The Bovine Genome Sequencing and Analysis Consortium, Elsik, C. G., Tellam, R. L., et Worley, K. C. (2009). The genome sequence of taurine cattle : A window to ruminant biology and evolution. *Science*, 324(5926) :522–528.

The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis : Multitissue gene regulation in humans. *Science*, 348(6235) :648–660.

Tollervey, D. (2004). Molecular biology : termination by Torpedo. *Nature*, 432 :456–457.

Treiber, T., Treiber, N., et Meister, G. (2018). Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. *Nature Reviews Molecular Cell Biology*.

Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P. E., Do, R., Mikkelsen, T. S., et Sankaran, V. G. (2016). Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*, 165 :1530–1545.

Vagner, S., Galy, B., et Pyronnet, S. (2001). Irresistible IRES. *EMBO reports*, 2 :893–898.

Van Damme, P., Hole, K., Pimenta-Marques, A., Helsens, K., Vandekerckhove, J., Martinho, R. G., Gevaert, K., et Arnesen, T. (2011). NatF contributes to an evolutionary shift in protein N-terminal acetylation and is important for normal chromosome segregation. *PLOS Genetics*, 7 :1–19.

Van den Steen, P., Rudd, P. M., Dwek, R. A., et Opdenakker, G. (1998). Concepts and principles of o-linked glycosylation. *Critical Reviews in Biochemistry and Molecular Biology*, 33 :151–208.

Van der Wijst, M., Brugge, H., de Vries, D., Deelen, P., Swertz, M., et Franke, L. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics*, 50 :493–497.

Vicens, Q., Paukstelis, P. J., Westhof, E., Lambowitz, A. M., et Cech, T. R. (2008). Toward predicting self-splicing and protein-facilitated splicing of group I introns. *RNA*, 14(10) :2013–2029.

Võsa, U., Esko, T., Kasela, S., et Annilo, T. (2015). Altered Gene Expression Associated with microRNA Binding Site Polymorphisms. *PLOS ONE*, 10(10) :1–24.

Wahid, F., Shehzad, A., Khan, T., et Kim, Y. Y. (2010). MicroRNAs : Synthesis, mechanism, function, and recent clinical trials. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1803 :1231 – 1243.

Wang, M., Hancock, T. P., Chamberlain, A. J., Vander Jagt, C. J., Pryce, J. E., Cocks, B. G., Goddard, M. E., et Hayes, B. J. (2018). Putative bovine topological association domains and CTCF binding motifs can reduce the search space for causative regulatory variants of complex traits. *BMC Genomics*, 19(1) :395.

Wang, X. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, 505 :117–120.

Weil, P. A., Luse, D. S., Segall, J., et Roeder, R. G. (1979). Selective and accurate initiation of transcription at the AD2 major late promotor in a soluble system dependent on purified RNA polymerase II and DNA. *Cell*, 18(2) :469 – 484.

Wen, J., Ladewig, E., Shenker, S., Mohammed, J., et Lai, E. C. (2015). Analysis of nearly one thousand mammalian mirtrons reveals novel features of dicer substrates. *PLOS Computational Biology*, 11 :1–29.

West, D. B., Pasumarthi, R. K., Baridon, B., Djan, E., Trainor, A., Griffey, S. M., Engelhard, E. K., Rapp, J., Li, B., Jong, P. J. d., et Lloyd, K. K. (2015). A lacZ reporter gene expression atlas for 313 adult KOMP mutant mouse lines. *Genome Research*, 25(4) :598–607.

West, S., Gromak, N., et J Proudfoot, N. (2004). Human 5' -> 3' exonuclease Xm2 promotes transcription termination at co-transcriptional cleavage sites. *Nature*, 432 :522–5.

Wickiser, J. K., Winkler, W. C., Breaker, R. R., et Crothers, D. M. (2005). The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Molecular Cell*, 18(1) :49 – 60.

Wierzbicki, A. T., Ream, T. S., Haag, J. R., et Pikaard, C. S. (2009). RNA Polymerase V transcription guides ARGONAUTE4 to chromatin. *Nature Genetics*, 41 :630–634.

Wilker, E. W., van Vugt, M. A., Artim, S., Huang, P., Petersen, C., Reinhardt, H., Feng, Y., A Sharp, P., Sonenberg, N., M White, F., et Yaffe, M. (2007). 14-3-3sigma controls mitotic translation to facilitate cytokinesis. *Nature*, 446 :329–32.

Wilkins, J. F. et Úbeda, F. (2011). Chapter 13 - diseases associated with genomic imprinting. In *Modifications of Nuclear DNA and its Regulatory Proteins*, volume 101 of *Progress in Molecular Biology and Translational Science*, pages 401 – 445. Academic Press.

Williams, R., Chan, E., Cowley, M. J., et Little, P. F. R. (2007). The influence of genetic variation on gene expression. *Genome research*, 17(12) :1707–16.

Willis, I. M. (1993). RNA polymerase III. *European Journal of Biochemistry*, 212 :1–11.

Wood, D. L. A., Nones, K., Steptoe, A., Christ, A., Harliwong, I., Newell, F., Bruxner, T. J. C., Miller, D., Cloonan, N., et Grimmond, S. M. (2015). Recommendations for accurate resolution of gene and isoform allele-specific expression in RNA-Seq data. *PLOS ONE*, 10(5) :1–27.

Xin, D., Hu, L., et Kong, X. (2008). Alternative promoters influence alternative splicing at the genomic level. *PLOS ONE*, 3(6) :1–8.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16 :15 – 24.

Yang, B., Damaschke, N., Yao, T., McCormick, J., Wagner, J., et Jarrard, D. (2016). Pyrosequencing for accurate imprinted allele expression analysis. *Journal of Cellular Biochemistry*, 116(7) :1165–1170.

Yang, B., Wagner, J., Yao, T., Damaschke, N., et Jarrard, D. F. (2013). Pyrosequencing for the rapid and efficient quantification of allele-specific expression. *Epigenetics*, 8(10) :1039–1042.

Yang, C., Li, X., Wang, Y., Zhao, L., et Chen, W. (2012). Long non-coding RNA UCA1 regulated cell cycle distribution via CREB through PI3-K dependent pathway in bladder carcinoma cells. *Gene*, 496 :8 – 16.

Yu, C.-Y., Chuang, C.-Y., et Kuo, H.-C. (2018). Trans-spliced long non-coding RNA : an emerging regulator of pluripotency. *Cellular and Molecular Life Sciences*, 75(18) :3339–3351.

Yuan, Y., H Wang, Z., et G Tang, J. (1999). Intra-a chain disulphide bond forms first during insulin precursor folding. *The Biochemical journal*, 343 Pt 1 :139–44.

Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., et Kinzler, K. W. (1997). Gene expression profiles in normal and cancer cells. *Science*, 276(5316) :1268–1272.

Zheng, G., Dahl, J., Niu, Y., Fedorcsák, P., Huang, C.-M., J Li, C., Vågbø, C., Shi, Y., Wang, W.-L., Song, S.-H., Lu, Z., P.G. Bosmans, R., Dai, Q., Hao, Y.-J., Xin, Y., Zhao, W., Tong, W.-M., Wang, X.-J., Bogdan, F., et He, C. (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Molecular cell*, 49 :18–29.

Zhuo, Z., Lamont, S. J., et Abasht, B. (2017). RNA-Seq analyses identify frequent allele specific expression and no evidence of genomic imprinting in specific embryonic tissues of chicken. *Scientific Reports*, 7.

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C. P., Sonstegard, T. S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J. A., et Salzberg, S. L. (2009). A whole-genome assembly of the domestic cow, bos taurus. *Genome Biology*, 10(4) :R42.

Zinshteyn, B. et Nishikura, K. (2009). Adenosine-to-inosine RNA editing. *Wiley Interdisciplinary Reviews : Systems Biology and Medicine*, 1(2) :202–209.

Zou, F., Chai, H. S., Younkin, C. S., Allen, M., Crook, J., Pankratz, V. S., et Ertekin-Taner, N. (2012). Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLOS Genetics*, 8(6) :e1002707.

)

Annexe A

SHORT TERM SCIENTIFIC MISSION
SCIENTIFIC REPORT

Mr Gabriel Guillocheau
INRA Jouy-en-Josas
Domaine de Vilvert
78350 Jouy-en-Josas
France

Reference COST Action CA15112
Host institution: The Roslin Institute, David Hume Research Group
Period: 01/01/2017 to 31/03/2017
Reference code: COST-STSM-ECOST-STSM-CA15112-010117-081649

Purpose of the STSM:

The aims of the Short Term Scientific Mission were to:

- Detect Single Nucleotide Variants (SNVs) showing Allele Specific Expression in additional ruminant species other than cattle (including, sheep and water buffalo);
- Strengthen collaborative links between INRA and Roslin;
- Facilitate sharing of data and analysis pipelines between FAANG institutions;
- Provide training in computational analysis of Allele Specific Expression (ASE) in the partner institution;
- Visit another European lab and improve my scientific network by interacting with other scientists from several different groups.

Description of the work carried out during the STSM:

During my mission at the Roslin Institute I applied a pipeline to detect Single Nucleotide Variants (SNVs) showing Allele Specific Expression (ASE) in sheep and water buffalo. I developed this bioinformatic pipeline during the first year of my PhD at INRA. Firstly, the pipeline detects variants by using GATK. Then, it filters results using ASEReadCounter and an N-masking method. Lastly, it predicts allelic imbalance by computing a binomial test.

For sheep, RNA-Seq data from bicep muscle (3 ewes and 3 rams) and the respective Whole Genome Sequencing (WGS) data generated by the Roslin Sheep Gene Expression Atlas Project were used. These animals were a cross breed produced by a Texel sire and a Scottish Blackface dam. We were particularly interested in investigating ASE in muscle as muscling and growth traits are of significant economic importance in sheep and I had already generated parallel results for

cattle bicep muscle with which we could compare the results. The pipeline detected 3,687,464 SNVs from the transcriptome and 27,608,126 SNVs from the genome. After using filters and statistical analyses, 18,721 ASE-SNVs were predicted.

For water buffalo, it was more complicated because of the lack of a good reference assembly. There is no chromosome information for the current water buffalo assembly but more than 300 thousand scaffolds. RNA-Seq data from longitudinal dorsal muscle (two male and two female Mediterranean buffalo) generated by the Roslin Water buffalo Atlas Project was used. To save time, we tried to perform the variant calling on the reference transcriptome, but it seemed less sensitive than using the genome. To compare these approaches, I relaunched the variant calling on the sheep bicep data using the transcriptome as a reference to estimate the impact of the reference used. SNVs predicted using transcriptome reference accounted for only 5% of those detected when using the genome reference, therefore the impact of using the transcriptome as a reference was too great. For the buffalo analysis I suggested removing small scaffolds from the genome to save time. The pipeline was edited to perform variant calling using filter on scaffold size.

At the end of my 3-month mission, I trained members of David Hume's group to use my pipeline. In addition during my time at Roslin I was able to provide computational help and support to a PhD student working on the water buffalo project facilitating analysis of this dataset and providing a useful collaborative link between the two laboratories. I've also written a comprehensive instruction manual for the pipeline. The pipeline I have developed will continue to be used to analyse sheep and water buffalo RNA-Seq datasets generated at Roslin and the results will be shared within the FAANG consortium.

Description of the main results obtained:

For sheep bicep muscle, my pipeline predicted 18,721 unique ASE-SNVs (See Table 1).

The majority of allelic imbalance phenomenon was predicted just in one individual. Interrogation of the dataset in more detail showed that there were just 7 ASE-SNVs common across 5 individuals and no ASE-SNVs predicted that were shared across all six individuals. Of the 7 ASE-SNVs shared across 5 individuals only 3 of these were in exonic regions.

Ind.	Predicted ASE	Het RNA SNVs	Het DNA SNVs	Common SNVs	Crossed ASE-SNVs
m1	26,536	360,444	8,747,923	308,219	5,391
m2	14,831	230,594	13,178,729	170,631	2,048
m3	25,739	405,506	12,224,251	173,484	2,829
f1	21,243	301,197	13,096,535	217,951	3,335
f2	22,839	317,388	13,164,705	144,888	2,454
f3	26,431	432,909	12,860,655	305,032	4,602

Table 1 Details of the number of SNVs detected for each individual sheep and the ASE-SNVs predicted on completion of the pipeline.

The analysis for sheep also indicated there was no link between the size of the chromosome and the number of SNVs showing ASE (See Figure 1).

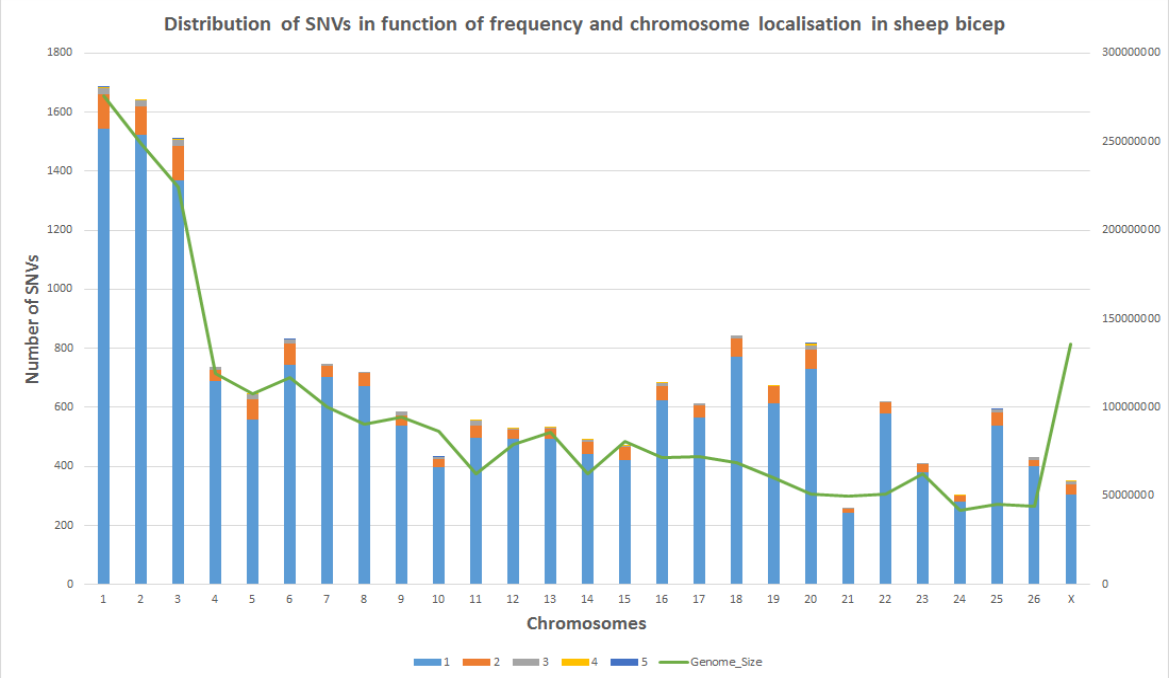

Figure 1 Distribution of SNVs in relation to chromosome in sheep bicep muscle.

In addition I compared the results generated with sheep bicep muscle with those I had previously generated for cattle, which showed a higher proportion of ASE-SNVs in exonic regions (synonymous, missense and coding sequence) in cattle relative to sheep and the reverse for intronic and intergenic regions (See Figure 2).

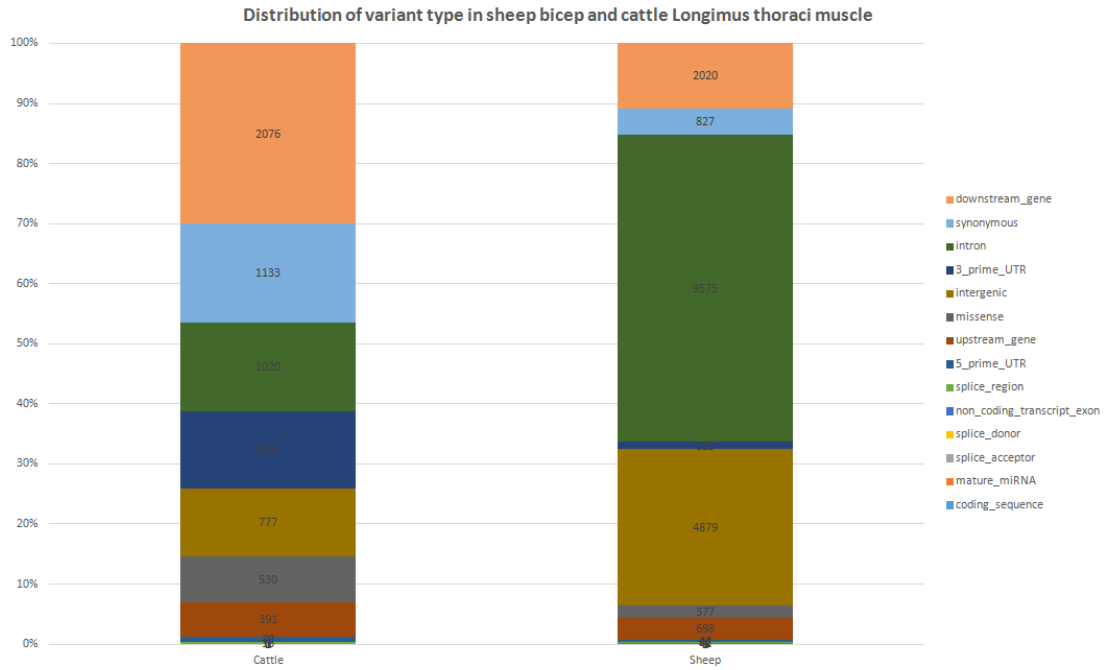

Figure 2 Distribution of variant type in sheep bicep and cattle longimus thoraci muscle.

For water buffalo, my pipeline predicted 5,098 unique ASE-SNVs from transcriptome (See Table 2). The prediction was done by using the reference transcriptome, so the lack of information is important.

The majority of allelic imbalance phenomenon was predicted just in one individual (3655 ASE-SNVs). Interrogation of the dataset in more detail showed that there were 311 ASE-SNVs common across 3 individuals and no ASE-SNVs predicted that were shared across all four individuals.

Ind.	Common ASE	Het RNA SNVs	Crossed ASE-SNVs
m1	5018	9131	1742
m3	4875	7018	1505
f1	4625	9295	1562
f3	5637	9863	2043

Table 2 Details of the number of SNVs detected for each individual buffalo and the ASE-SNVs predicted on completion of the pipeline.

Future collaboration with the host institution and conclusions:

During my time at the Roslin Institute I was able to meet other researchers within and outwith the Hume group with shared research interests, including analysis of ASE. This was very useful experience for me as an early career scientist and I will keep in contact and hope to collaborate with them for future projects. I hope to continue my analysis of the sheep and water buffalo datasets, particularly of muscle tissue, and will continue to collaborate with members of the Hume Group to do this and also contribute to publication of the work. My time at Roslin was very valuable in introducing me to new people, with shared research interests, and allowing me to develop my communication and data analysis skills. I have helped to build computational resources, in the form of an ASE analysis pipeline, at Roslin and facilitated training of post docs and PhD students in its use. In return I have had access to datasets from additional ruminant species that I can compare with the dataset I am working on from cattle and will be included in the publication of ASE analysis for these additional species. It has been a very productive experience and the links between INRA and Roslin, two key contributors to the FAANG consortium, are stronger and more productive as a result.

Publications, Communications et Encadrement

Posters

ISAG 2016 (Salt Lake City, Etats-Unis) : Study of polymorphisms modifying gene expression regulation in cattle ; G. M. Guillocheau et D. Rocha.
Cambridge Livestock Genomics Meeting (Cambridge, Royaume-Unis) : Identification of polymorphisms modifying gene expression regulation in cattle ; G. M. Guillocheau et D. Rocha.
ISAG 2017 (Dublin, Irlande) : Identification of polymorphisms modifying gene expression regulation in cattle ; G. M. Guillocheau et D. Rocha.
Journée de Doctorants ABIES 2017 (Paris, France) : Identification of polymorphisms modifying gene expression regulation in cattle ; G. M. Guillocheau et D. Rocha.

Communications orales

Séminaire de Doctorants du Département de Génétique Animale (Bruz, France) : Identification of polymorphisms modifying gene expression regulation in cattle ; G. M. Guillocheau et D. Rocha. Présentation de 10 minutes.
ISAG 2017 (Dublin, Irlande) : Identification of polymorphisms modifying gene expression regulation in cattle ; G. M. Guillocheau et D. Rocha. Présentation de 15 minutes.

Publications

Genetic variability of the activity of bidirectional promoters : a pilot study in bovine muscle : C. Meersseman, R. Letaief, V. Léjard, E. Rebours, G. Guillocheau, D. Esquerré, D. Rocha ; DNA Research 2017.
Survey of allele specific expression in bovine muscle : G. M. Guillocheau, A. El Hou, C. Meersseman, D. Esquerré, E. Rebours, R. Letaief, M. Simao, N. Hypolite, E. Bourneuf, N. Bruneau, A. Vaiman, C. J. Vander Jagt, A. J. Chamberlain, and D. Rocha. Soumis à Scientific Report le le 5 juillet, révision majeure le 17 Octobre.

Stages encadrés

Majda Arif (M1 en bioinformatique), stage de 8 semaines sur "Analyse de l'empreinte parentale chez le bovin".

Abdelmajid El Hou (M2 européen PRIAM), stage de 6 mois sur "Analyse de polymorphismes altérant la régulation de l'expression des gènes, chez le bovin".

Doruntine Fezjovski (L2 en bioinformatique), stage de 8 semaines sur "Identification dans le génome de la souris et du bovin des sites cibles de Cas".

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Etude des polymorphismes altérant la régulation de l'expression des gènes chez le bovin.

Mots clés : bio-informatique, séquençage, transcriptomique, polymorphismes, promoteur

Résumé : Un nombre croissant de gènes et régions génomiques sont associés à des pathologies ou des phénotypes d'intérêt, soit par analyse de liaison ou analyse d'association. Il est crucial d'arriver à identifier les variants génétiques causaux. Les régulateurs ayant l'effet le plus important sont le plus souvent des polymorphismes régulateurs exerçant un effet en *cis*, près des gènes pour lesquels le niveau d'expression est altéré. L'objectif global de la thèse est d'identifier à grande échelle les polymorphismes chez la vache qui potentiellement altèrent la régulation de l'expression des gènes et affectent des phénotypes d'intérêt.

Nous avons développé une approche pour déterminer les SNPs (Single Nucleotide Polymorphisms) causant ou étant impliqué dans une régulation de l'expression des gènes. Dans cet objectif, nous avons analysé chez 19 taurillons Limousin le génome et le transcriptome musculaire et chez 6 vaches Holstein le génome et le transcriptome de 8 tissues dont utérus et ovaire. Chez les mâles Limousin, nous avons identifié 5 658 SNPs montrant une expression allèle spécifique (ASE-SNPs) dans 13% des gènes exprimés dans le muscle et avons lié certains d'entre eux à des SNPs dans une région régulatrice. On a aussi identifié des gènes d'intérêt liés à la qualité de la viande (*AOX1*, *PALLD* et *CAST*) qui présente un déséquilibre allélique. Chez les femelles Holstein, nous avons identifié 33 534 ASE-SNPs dans les 8 tissus dont 3 369 ASE-SNPs pour les données de muscle, 5 771 pour les données d'ovaire et 5 499 pour les données d'utérus. L'analyse de ces deux jeux de données bovins a permis une nouvelle cartographie des gènes soumis à ASE. Il s'agit de la première analyse de cette ampleur pour la race Limousine.

Les résultats de ces études permettent d'approfondir la compréhension de la régulation de l'expression des gènes chez le bovin, notamment en identifiant des polymorphismes causaux candidats et en apportant des nouvelles méthodes pour les détecter.

Title : Study of polymorphisms modifying gene expression regulation in cattle.

Keywords : bioinformatics, sequencing, transcriptomics, polymorphisms, promoter

Abstract : An increasing number of genes and genomic loci have been associated with diseases or phenotypes of interest, by either linkage or association studies. Identifying causative genetic variants is crucial. The regulators with the strongest effect tend to be *cis*-acting regulatory polymorphisms, close to genes for which altered mRNA expression was detected. The overall objective of this PhD project is to develop a large-scale approach to identify regulatory polymorphisms that potentially alter the regulation of gene expression and impact phenotypes of interest, in cattle.

We have developed an approach to ascertain causative SNPs (Single Nucleotide Polymorphisms) of gene expression regulation. To this end, we analysed genome and muscle transcriptome from 19 Limousine bull calves and genome and transcriptome of eight tissues (including ovary and uterus) transcriptome from 6 Holstein cows. For the Limousine breed, we identified 5,658 SNPs showing an allele-specific expression (ASE-SNPs) in 13% of genes with detectable expression in muscle; we linked some of them to SNPs in a regulatory region. Interestingly, we found genes involved in meat quality traits (*AOX1*, *PALLD* and *CAST*) with an allelic imbalance. For the Holstein breed, we identified 33,534 ASE-SNPs across 8 tissues including 3,369 ASE-SNPs from muscle data, 5,771 from ovary data and 5,499 from uterus data. By analysing these two data records, we discovered genes impacted by ASE. This study is the first done for the Limousine breed and the second for thr Holstein breed.

The results of these studies provide a best understanding of gene expression regulation in cattle, in particular by identifying candidate causal polymorphisms and by proposing new methods to detect them.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France