HAL open science

Etude des patrons de recombinaison, de leur déterminisme génétique et de leurs impacts en sélection génomique

Morgane Petit

▶ To cite this version:

Morgane Petit. Etude des patrons de recombinaison, de leur déterminisme génétique et de leurs impacts en sélection génomique. Sciences agricoles. Institut National Polytechnique de Toulouse - INPT, 2017. Français. NNT : 2017INPT0083 . tel-04228134v2

HAL Id: tel-04228134

https://theses.hal.science/tel-04228134v2

Submitted on 4 Oct 2023

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

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

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (INP Toulouse)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

Mme MORGANE PETIT

le mardi 17 octobre 2017

Titre :

Etude des patrons de recombinaison, de leur déterminisme génétique et de leurs impacts en sélection génomique

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GenPhySE)

Directeur(s) de Thèse :

MME CAROLE MORENO

M. BERTRAND SERVIN

Rapporteurs :

M. DIDIER BOICHARD, INRA JOUY EN JOSAS

M. MATTHIEU FALQUE, INRA PARIS

Mme CAROLE CHARLIER, UNIVERSITE DE LIEGE

Membre(s) du jury :

M. DIDIER BOICHARD, INRA JOUY EN JOSAS, Président

M. ALAIN VIGNAL, INRA TOULOUSE, Membre

Etude des patrons de recombinaison, de leur déterminisme génétique et de leurs impacts en sélection génomique en race ovine Lacaune

Morgane PETIT

Octobre 2017

Remerciements

Je tiens tout d'abord à remercier les membres de mon jury de thèse : Didier Boichard, Carole Charlier, Alain Ducos, Matthieu Falque et Alain Vignal pour avoir accepté de lire, corriger et évaluer cette thèse. Je tiens également à remercier tout particulièrement les membres de mon comité de thèse : Tom Druet, Laurent Duret, Alain Pinton et Pierre Sourdille, pour leurs conseils et leur aide tout au long de ces trois années de doctorat. Je remercie également la région Occitanie, ainsi que le Métaprogramme Selgen pour avoir financé ces travaux.

Je remercie bien sûr mes deux encadrants, mes deux « parents » de thèse : Carole Moreno et Bertrand Servin, sans qui cette thèse n'aurait pas pu être possible. Merci à vous de m'avoir aidée à avancer au cours de ce doctorat, de vous être autant impliqués et de m'avoir permis d'obtenir tous les résultats que je présente aujourd'hui et surtout merci ! pour votre patience et votre compréhension, malgré tous mes doutes, mes questionnements, mes baisses de moral et mon manque de confiance en moi. Merci pour tout !

Je remercie également Jean-Michel Astruc et les Organismes de Sélection Lacaune et Romane. S'ils n'existaient pas, il n'y aurait pas de données, donc merci de les mettre à disposition et de nous permettre de les utiliser et de les étudier. Merci aussi à Julien Sarry, Stéphane Fabre et Laurence Drouilhet qui ont réalisé les génotypages et qui ont découvert les mutations. Heureusement que vous étiez là, parce que les pipettes et moi, cela fait 12 ! Merci aussi à Hélène Larroque pour m'avoir permis d'étudier l'imputation et les puces basse densité.

Un doctorat, cependant, ce n'est pas seulement un doctorant et ses encadrants, c'est aussi toute une équipe de soutien au quotidien, donc là je dois remercier toute ma « team » qui m'a tant soutenue (voire supportée) tout au long de ces trois (virgule 5) années. Je tiens d'ailleurs à m'excuser auprès de vous pour la fin de cette thèse… Je commencerai avec Marjo', après tout c'est avec elle que tout a commencé ! Merci à toi de m'avoir accueillie (plus ou moins forcée, certes) le premier jour, de m'avoir montré le chemin de la salle de réunion, de la cantine et tout simplement du bureau, parce que franchement, si j'ai une chose à suggérer pour les travaux, mettez des pancartes dans les couloirs pour les pauvres stagiaires ! Et finalement, merci Marjo' de ne pas m'avoir lâché la main à partir de cet instant l Les M&M's comme on nous a appelé : mots

croisés et fous rires du vendredi, puis le bâtiment C et retour dans le bâtiment E, plus dans le même bureau, mais finalement pas très loin ! Bref, mon premier soutien ! Merci pour ton aide, pour ta patience, pour tes conseils, pour m'avoir remonté le moral et pour t'être occupée de ma forme physique (tu es la seule qui a réussi à me refaire courir !). Ne t'inquiète pas, je ne t'oublie pas Charlotte ! J'ai commencé à faire mieux ta connaissance grâce à cette très belle visite de La Fage (les seuls moutons que j'aurai vu en vrai !), visite, que paraît-il je peux m'attribuer le mérite en plus ! Merci à toi d'être toujours de si bonne humeur, si souriante et si pleine de couleurs ! Te voir chaque jour, pleine de vie, d'anecdotes et de sourires a été un plaisir et un remède efficace contre les jours « gris » ! J'adore discuter et rire avec toi, suivre tes aventures et t'accompagner chez le véto' ! J'en viens ensuite à ma « CléClé », ex-copine de Gym Direct (oui, tu m'as lâchement abandonné). Que dire ? Reste toi-même, ne change rien, je t'aime comme tu es ! J'adore ta spontanéité, tes remarques drôles malgré toi (bien que parfois, j'espère que tu le fais exprès) ! Si tu pars en Corée, tu vas vraiment nous manquer, mais en même temps, tu apporteras aux Coréens tant de fous rires, qu'il serait dommage de ne pas te partager, mais reviens-nous quand même ! Merci aussi à Jojo ! Tu m'as tant soutenu lors de mon séjour au bâtiment C ! Les repas de midi n'auraient pas été les mêmes sans toi, sans parler de toutes nos soirées « filles » (+ Maël), des potinages (ah oui, le meilleur à l'INRA, ce sont le potins !) et de toutes les activités dans lesquelles tu m'as accompagnée ! Merci à Mathou' ! Avec toi, on a échangé nos petits tracas de thèse, nos plaintes, mais aussi nos succès ! Tu es une fille pleine d'humour, toujours super bien habillée (« Ma Chéwiiii ») et on te regrette à Paris, mais j'y crois, tu vas revenir vers nous ! Merci aussi à Noémie, toujours si souriante ! Il faut que tu nous donnes ton secret ! Tu m'as aussi beaucoup apporté et j'adore quand tu prends notre défense à nous, pauvres petits doctorants ! Merci aussi à Emilie, ma première « stagiaire » que j'ai si bien encadrée (n'est-ce pas ?). Merci pour ton aide avec *PRDM9* et merci de ton soutien tout au long de l'été dernier et merci aussi pour le petit cadeau de fin de stage, c'était vraiment trop mignon ! Et je n'oublie pas mes copines d'Amour est dans le Pré ! Le lundi soir est le meilleur moment de la semaine ! Merci Héloïse, Valérie et Sophie : #Stagiaire (Et mention spéciale à Héloïse pour m'avoir fait découvrir l'Eurovision) ! Et Sophie, n'oublie pas tout ce qu'on t'a transmis ! Merci à Pauline, je croise très fort les doigts pour ton concours ! Merci de m'avoir si bien intégrée à tous les « jeunes de GenPhySE » ! Merci aussi à Claire, Marc, Mathieu, Jason, Cyriel, Maxime, Céline et Sophie Brard

pour votre bonne humeur au quotidien et pour agrémenter les pauses café et la cantine ! Mention spéciale à Sophie pour son aide pour mon CV !

J'aimerais aussi remercier mes deux équipes d'accueil : GesPR et Dynagen. Merci de votre accueil, de votre aide pour les répétitions, pour les conseils et les discussions sur les résultats présentés. Merci pour le super repas GesPR ! Je remercie également toute l'unité GenPhySE et, bien qu'on se plaigne parfois, nous avons bien conscience que nous sommes chouchoutés chez GenPhySE ! Merci Nancy, Florence, Manuela, Viviane et Evelyne pour votre aide lors des congrès et notamment lors des avions annulés… Et merci Yann pour l'aide sur R et, tu vois, je l'ai réussie ma thèse ! Je t'offrirai un dernier paquet de M&M's pour que tu penses à moi ! Et merci Hélène Gilbert de m'avoir acceptée dans ton bureau et d'être si sympathique !

Enfin, je remercie ma famille pour m'avoir toujours encouragée dans mes études et m'avoir poussée à me dépasser, certes je ne vous ai pas trop embêté avec la thèse, mais vous avez déjà eu votre lot entre les contrôles de maths, le Bac et la prépa'… ! Et merci Mon Dou, merci d'être toujours là, d'être si gentil, si patient, surtout que c'est particulièrement toi qui a dû supporter mes sautes d'humeur, mes crises de déprime, d'angoisse et de stress… Tu as toujours été présent, aimant et tu m'as toujours encouragée. Alors, vraiment un grand merci et surtout ne change jamais…

Rêve ta vie en couleurs, c'est le secret du bonheur…

Table des Matières

Introduction 3

Chapitre 1 : La recombinaison méiotique ; biologie fonctionnelle, phénotypes étudiés et déterminisme génétique associé 7

I. Biologie fonctionnelle de la recombinaison 7

I. 1. La recombinaison méiotique : définition et rôle 7

I. 2. La recombinaison génétique a lieu au cours de la méiose 8

I. 3. Détails de la Prophase I : étape essentielle de la recombinaison méiotique 9

I. 4. La formation des crossing-overs 13

II. Etude de la recombinaison méiotique 22

II. 1. Méthodes d'estimation des cartes de recombinaison 22

II. 2. Etude des cartes de recombinaison au niveau génomique 37

II. 3. Etude des cartes de recombinaison au niveau local 41

III. Variation inter-individuelle de la recombinaison 48

III. 1. Le taux global de recombinaison 48

III. 2. La localisation de la recombinaison 50

IV. Le déterminisme génétique de la variation inter-individuelle du taux de recombinaison et de la variation inter-individuelle de la localisation de la recombinaison 54

IV. 1. Méthodes de détection des QTLs 54

IV. 2. Le déterminisme génétique de la variation inter-individuelle du taux de recombinaison 64

IV. 3. Le déterminisme génétique de la variation inter-individuelle de la localisation de la recombinaison 67

IV. 4. Etat de l'art chez le mouton 70

V. Objectifs de la thèse 71

Chapitre 2 : Utilisation de deux jeux de données différents pour créer des cartes génétiques de haute résolution 75

I. Deux jeux de données disponibles en Lacaune 75

I. 1. La race ovine Lacaune lait : sujet de l'étude 75

I. 2. Un grand pédigrée génotypé avec une puce moyenne densité 78

I. 3. Un échantillon d'animaux non apparentés génotypés avec une puce haute densité 79

II. L'étude des cartes de recombinaison en données familiales 80

II. 1. La détection des crossing-overs 80

II. 2. L'estimation des taux de recombinaison 82

II. 3. Les Lacaune ont des patrons de recombinaison communs aux animaux d'élevage 84

III. L'étude des cartes de recombinaison en données populationnelles 88

II. 4. Méthode de détection des points chauds de recombinaison 88

II. 5. Les points chauds de recombinaison existent en Lacaune 91

III. La création de cartes génétiques de haute résolution grâce à la combinaison des deux jeux de données 94

III. 1. Méthode d'obtention des cartes de haute résolution 94

III. 2. Impact des points chauds sur le taux de recombinaison 96

III. 3. Détection de signatures de sélection 98

IV. Comparaison des cartes génétiques entre Lacaune et Soay 103

V. Discussion 107

V. 1. La détection des points chauds de recombinaison 107

V. 2. La combinaison des deux jeux de données populationnel et familial 109

VI. Le biais d'usage des points chauds : autres pistes d'étude 111

VI. 1. Essais d'une méthode indirecte pour détecter ce phénotype 111

VI. 2. Recherche du gène PRDM9 à l'aide de motifs d'ADN spécifiques 112

VI. 3. Conclusion intermédiaire : la recherche de PRDM9 128

Chapitre 3 : Etude du déterminisme génétique de la variation intert individuelle du taux de recombinaison 131

I. Observation du phénotype « taux de recombinaison individuel » 131

II. La détection de *QTLs* suite à l'utilisation d'une *GWAS* 134

II. 1. Amélioration de la densité en marqueurs grâce à l'imputation 134

II. 2. La GWAS a permis d'identifier 3 QTLs 137

II. 3. Recherche de mutations candidates dans le gène RNF212 148

III. Le déterminisme génétique du taux de recombinaison individuel diffère entre les Soay et les Lacaune mâles 151

IV. Conclusions et perspectives 153

IV. 1. Estimation du taux de recombinaison comme un index 153

IV. 2. Comparaison des QTLs avec les autres espèces 154

IV. 3. La recherche de mutations causales 157

IV. 4. Nouvelle approche pour l'étude du déterminisme grâce à l'utilisation d'une race croisée, la Romane 158

Chapitre 4 : Utilisation des cartes génétiques en sélection : création de puces basse densité 163

Comparison of imputation accuracy using *SNPs* sets based on a physical map or on a genetic map 164

Abstract 164

Introduction 165

Materials and Methods 166

Study population 166

Creation of low-density SNPs sets 166

Imputation of the low-density SNPs sets on the 50K SNPs array 168

Test of the imputation quality for the 8 low-density SNPs sets 168

Results 169

Test of the imputation quality for the 8 low-density SNPs sets 169

Discussion 179

Conclusion 179

Literature cited 180

Discussion générale 183

I. Amélioration des cartes de recombinaison 184

I. 1. Amélioration des cartes de recombinaison méiotique 184

I. 2. Création de cartes de recombinaison haute résolution 186

II. Préciser le déterminisme génétique 189

II. 1. Le taux de recombinaison individuel 189

II. 2. Le biais d'usage des points chauds 190

III. Intérêts de l'étude de la recombinaison en sélection 193

III. 1. Utilisation de la recombinaison pour la création de nouvelles puces 193

III. 2. Utilisation du taux de recombinaison pour augmenter la réponse à la sélection 193

III. 3. La recombinaison accélère l'introgression génique 197

III. 4. Conclusions 198

Conclusion générale 202

Bibliographie 206

Annexes 222

Abstract 223

Introduction 223

Materials and Methods 227

Study Population and Genotype Data 227

Recombination Maps 227

Meiotic recombination maps from pedigree data 227

Historical recombination maps from the diversity data 229

Combination of meiotic and historical recombination rates and construction of a high resolution recombination map 230

Comparison with Soay sheep recombination maps and integration of the two datasets to produce new male recombination maps in Sheep 232

Genome-Wide Association Study on Recombination Phenotypes 233

Genome-wide Recombination Rate (GRR) 233

Genotype Imputation 233

Single- and multi-QTLs GWAS on GRR 234

Variant Discovery and Additional Genotyping in *RNF212* 235

Identification and assignation of the RNF212 sheep genome sequence 235

Variant discovery in RNF212 in the Lacaune population 236

Genotyping of mutations in RNF212 236

Results 237

High-Resolution Recombination Maps 237

Meiotic recombination maps: genome-wide recombination patterns 237
Estimation of historical recombination rates and identification of crossover hotspots 238
High-resolution recombination maps combining family and population data 239
Improved male recombination maps by combining Lacaune and Soay sheep data 241
Genetic Determinism of Genome-wide Recombination Rate in Lacaune sheep 242
Genetic and environmental effects on GRR 242
Genome-wide association study identifies three major loci affecting GRR in Lacaune sheep... 243
Mutations in the RNF212 gene are strongly associated to Genome-wide Recombination Rate variation in Lacaune sheep 246
The genetic determinism of recombination differ between Soay and Lacaune males 247
Discussion 248
Fine-scale Recombination Maps 248
Determinism of Recombination Rate in sheep populations 251
Acknowledgments 253
Literature Cited 254
Contexte et état de l'art 259
Objectifs 260
Caractère novateur ou exploratoire du projet 261
Plan de travail et calendrier 261
Dispositif Expérimental 261
Calendrier 262
Résultats déjà acquis et résultats complémentaires attendus 262
Bibliographie 263
Coop G, Wen Z, Ober C, Pritchard JK, Przeworksi M. (2008) High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans Science 31 Jan 2008 319(5868):1395-8 263
Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. (2005) A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science 14 Oct 2005 : 321-324 263
Justification budgétaire 263

Liste des Publications

Revues internationales avec comité de lecture :

Soumis ou en préparation :

1. **Petit M., Astruc JM., Sarry J., Drouilhet L., Fabre S., Moreno C. et Servin B.** (2017). Variation in recombination rate and its genetic determinism in sheep (*Ovis Aries*) populations from combining multiple genome-wide datasets. *BioRxiv*. (Chapitres 2 et 3).
2. **Petit M., Astruc JM., Chassier M., Larroque H., Servin B. et Moreno C.** Comparison of imputation accuracy using *SNPs* sets based on a physical map or a genetic map. *En préparation*. (Chapitre 4).

Liste des Communications

Communications orales

Avec comité de lecture

1. **Petit M.,** Moreno C. et Servin B. Fine-scale recombination maps in the Sheep. *Plant & Animal Genome Conference*, San Diego, Etats-Unis, 8 – 13 Janvier 2016.
2. **Petit M.,** Fabre S., Sarry J., Moreno C. et Servin B. Variation in recombination rate and its genetic determinism in sheep (*Ovis Aries*) populations from combining multiple genome-wide datasets. *European Federation of Animal Science*, Tallin, Estonie, 28 Août – 1er Septembre 2017.

Séminaires/workshops/groupes de travail

1. **Petit M.,** Moreno C. et Servin B. Etude des patrons de recombinaison, de leur déterminisme génétique et de leur impact sur l'efficacité de la sélection génomique chez les ovins. *Comité de Pilotage du Comité National des Brebis Laitières*, Toulouse, France, 15 Octobre 2015.
2. **Petit M.,** Moreno C. et Servin B. Fine-scale recombination maps in the Sheep. *19ème Séminaire des Doctorants du Département de Génétique Animale*, Toulouse, France, 16 – 17 Mars 2016.

Posters

1. **Petit M.,** Moreno C. et Servin B. Fine-scale recombination maps in the Sheep. *Quantitative Genetics & Genomics*, *Gordon Research Conference*, Pise, Italie, 22 – 27 Février 2015.
2. **Petit M.,** Moreno CR. et Servin B. Fine-scale recombination mapping in the Sheep. *18ème Séminaire des Doctorants du Département de Génétique Animale*, La Rochelle, France, 21 – 22 Mai 2015.
3. **Petit M.,** Moreno CR. et Servin B. Fine-scale recombination mapping in the Sheep. *1er Séminaire des Doctorants du Métaprogramme Selgen*, Paris, France, 26 Mai 2016.

4. **Petit M.,** Moreno CR. et Servin B. Study of recombination rate in Lacaune sheep informs its evolution in Soay sheep. *Quantitative Genetics & Genomics*, *Gordon Research Conference*, Houston, Etats-Unis, 26 Février – 3 Mars 2017.

Lexique des Abréviations

ADN : Acide Désoxyribonucléique

BF : Bayes Factor

BGC : Biaised Gene Conversion

BOGM : Based on Genetic Map

BOPM : Based on Physical Map

BSLMM : Bayesian Sparse Linear Mixed Model

ChIP-sequencing : Chromatin Immuno-Precipitation

CLO : Contrôle Laitier Officiel

CLS : Contrôle Laitier Simplifié

CO : Crossing-Over

CR : Concordance Rate

CS : Complexe Synaptonémal

DAPI : *DNA* fluorochrome 4',6-diamidino-2-phenylindol

DL : Déséquilibre de Liaison

DSB : Double-Strand Break

EN : Early Nodule

EM : Expectaction-Maximization

ES : Entreprise de Sélection

FAO : Food and Agricultural Organization of the United Nations

FDR : False Discovery Rate

FISH : Fluorescent *in situ* Hybridization

GWAS : Genome-Wide Association Studies

HIA : Haematoxylin-Iron Alum

HMM : Hidden Markov Model

iHS : Integrated Haplotype Score

INRA : Institut National de la Recherche Agronomique

LINE : Long Interspersed Nuclear Elements

LLR : Log Likelihood Ratio

LN : Late Nodule

LOD : Lod-Score

LRT : Likelihood-Ratio Test

LTR : Long Terminal Repeat sequence

MAF : Minor Allele Frequencies

MCMC : Méthode de Monte-Carlo par chaînes de Markov

MHC : Major Histocompatibility Complex

NCO : Non-Crossing-Over

OPA : Obtention des Paternités par Assignation

OS : Organisme de Sélection

PAR : Pseudo-Autosomale Region

PCR : Polymerase Chain Reaction

PIP : Probabilité Postérieure d'Inclusion

PGE : Part de variance Génétique Expliquée

PWM : Position Weight Matrix

QTL : Quantitative Trait Locus

QTV : Quantitative Trait Variant

RFLP : Restriction Fragment Length Polymorphism

SDSA : Synthesis-Dependent Strand-Annealing

SNP : Single Nucleotide Polymorphism

SSLP : Simple Sequence Length Polymorphism

SVM : Support Vector Machine

TSS : Tanscription Start Sites

Liste des Figures

Figure 1 : Différentes phases de la méiose en photographie (gauche) et en schéma (droite), d'après Science Notebook : http://tmsseventhgradeteam.weebly.com/notebook/archives/03-2016). 9

Figure 2 : Les différentes étapes de la Prophase I (D'après Baudat et al., 2013) 13

Figure 3 : Mécanisme de formation des différents évènements de recombinaison (d'après Saintenac, 2012) 16

Figure 4: Identification des crossing-overs (d'après Chowdhury et al., 2009) 29

Figure 5 : Répartition du taux de recombinaison dans différentes espèces 40

Figure 6: Schématisation de l'imputation 63

Figure 7 : Comparaison entre la sélection classique et la sélection génomique (d'après Moreno et Sallé, 2011) 77

Figure 8 : Répartition des 5 940 Lacaune dans deux pédigrées 79

Figure 9 : Nombre de crossing-overs obtenus dans les deux pédigrées 81

Figure 10 : Patrons de recombinaison le long des autosomes des Lacaune 85

Figure 11 : Taux de recombinaison moyen de chaque autosome en Lacaune 87

Figure 12 : Observation de la distribution de la recombinaison populationnelle de base le long du génome 90

*Figure 13 : Distribution des intensités de recombinaison λ_j sur les intervalles **j** de la 600K* 91

Figure 14 : Distribution de la recombinaison sur le génome 92

Figure 15 : Représentation de la distribution sur le génome à l'aide du coefficient de Gini ... 93

Figure 16 : Effet des points chauds sur le taux de recombinaison familial 97

Figure 17 : Taux de recombinaison populationnel et familial dans des fenêtres de 1 Mb 98

Figure 18 : Impact des signatures de sélection sur le taux de recombinaison populationnel .. 99

Figure 19 : Intensité relative du taux de recombinaison populationnel par rapport au taux de recombinaison familial dans des fenêtres de 1 Mb 100

Figure 20 : Arbre phylogénétique des différentes races de mouton (d'après Rochus et al., 2017) 105

Figure 21 : Comparaison des taux de recombinaison entre les Lacaune et les Soay 106

Figure 22 : Schéma d'un doigt de zinc (Klug, 2010) 113

Figure 23 : Séquence de PRDM9 de Ovis aries issue de Ensembl 115

Figure 24 : Observation des séquences de l'exon 10 du mouton, de la vache et de l'Homme 116

Figure 25 : « Scaffold » JH922946 de Ovis aries avec 10 exons annotés 117

Figure 26 : Comparaison entre le « scaffold » et la fin du chromosome 1 118

Figure 27 : Alignement multiple de l'exon 10 de PRDM9 entre le mouton, la vache, l'Homme et la chèvre 120

Figure 28 : Logo obtenu avec la séquence protéique et la détection des doigts de zinc 123

Figure 29 : Histogramme représentant la distribution des scores des motifs 125

Figure 30 : Diagramme quantile-quantile des scores normalisés obtenus en comparant un point chaud et un point froid 126

Figure 31 : Variation individuelle du taux de recombinaison parmi les Lacaune mâles 132

Figure 32 : Effet du mois d'insémination sur le nombre moyen de crossing-overs par méiose. 133

Figure 33 : Identification d'un pic avec une GWAS sur 50 000 marqueurs 135

Figure 34 : Validation de l'imputation 137

Figure 35 : Identification de 2 QTLs majeurs grâce à une GWAS 138

Figure 36 : Confirmation de 2 QTLs majeurs 139

Figure 37 : Zoom sur le résultat de GWAS pour le chromosome 7 142

Figure 38 : Alignement local des génomes ovins et humains à proximité du QTL du chromosome 7 144

Figure 39 : Zoom sur le résultat de GWAS pour le chromosome 6 145

Figure 40 : Structure du gène RNF212 dans des espèces variées 147

Figure 41 : Déséquilibre de liaison entre les polymorphismes du gène RNF212 et les SNPs du

QTL du chromosome 6 150

Figure 42 : Comparaison des résultats de GWAS pour le chromosome 6 entre les Lacaune mâles (graphe du haut), les Soay mâles (graphe du milieu) et les Soay femelles (graphe du bas) 152

Figure 43 : Distribution de la taille des intervalles de résolution des crossing-overs détectés dans 3 dispositifs. 159

Figure 44 : Reconstruction des origines populationnelles 160

Liste des Tableaux

Tableau 1 : Récapitulatif des protéines intervenant dans le processus de la recombinaison .. 21

Tableau 2 : Gènes candidats proposés pour la variation inter-individuelle de la recombinaison 64

Tableau 3 : Régions du génome où les taux de recombinaison populationnel et familial sont significativement différents 102

Tableau 4 : Comparaison des données disponibles en Soay mâles et en Lacaune mâles 104

Tableau 5 : Diversité des séquences des doigts de zinc dans la séquence protéique de l'exon 10 de PRDM9 présent sur le « scaffold » JH92946 chez le mouton 122

Tableau 6 : Décomposition de la variation du taux de recombinaison individuel chez les mâles Lacaune 134

Tableau 7 : SNPs statistiquement associés avec le phénotype du taux de recombinaison individuel 140

Tableau 8 : Mutations détectées dans le gène RNF212 149

Introduction Générale

Introduction

La recombinaison méiotique est un processus biologique fondamental, nécessaire au maintien de la diversité génétique et à l'évolution des génomes eucaryotes (Baudat *et al.*, 2013). Il s'agit d'un mécanisme universel présent aussi bien dans le règne animal que dans le règne végétal. La recombinaison méiotique a donc été décrite dans plusieurs espèces animales, notamment chez certains animaux d'élevage, ainsi que chez plusieurs plantes.

L'étude de la recombinaison peut permettre de mieux comprendre ses mécanismes d'action et ses rôles biologiques. Elle peut également avoir un intérêt plus appliqué, notamment en sélection génomique. Cette sélection a pour but d'améliorer les index génétiques des jeunes animaux, ce qui permet de les sélectionner dès leurs premiers mois de vie (Duchemin *et al.*, 2012). De plus, la sélection génomique conduit à une réduction des intervalles de génération et à une possible amélioration de l'intensité de sélection. Elle est très utile pour les organismes de sélection car elle permet de réduire les coûts des schémas de sélection, de réduire le temps de transmission du gain génétique entre le noyau de sélection et les troupeaux commerciaux, d'augmenter le progrès génétique et d'inclure de nouveaux caractères dans les schémas de sélection (Duchemin *et al.*, 2013).

Pour que la sélection génomique puisse fonctionner dans une race, il est nécessaire de connaître le déséquilibre de liaison au sein de la population étudiée, et plus ce déséquilibre est fort, plus la sélection génomique pourra être efficace (Rupp *et al.*, 2016). Caractériser les taux de recombinaison pourrait ainsi amener une meilleure connaissance de ce déséquilibre de liaison ; en effet lorsque le taux de recombinaison est important, le déséquilibre de liaison entre les marqueurs est faible et donc meilleure sera la sélection génomique. De plus, des taux de recombinaison élevés pourraient également permettre d'augmenter la réponse à la sélection et de maintenir la variation génétique dans une population sous sélection, en augmentant la variation génétique parmi les individus sélectionnés (Battagin *et al.*, 2016). Utiliser la recombinaison génétique pourrait également permettre d'accélérer l'introgression génique.

La race Lacaune est la première race ovine française à bénéficier de la mise en place d'un schéma génomique. En effet, il existe de nombreuses données de génotypage ; l'*ADN* des béliers ayant été conservé depuis 1995. Pourtant, ces données n'avaient pas encore été exploitées pour

la description de la recombinaison génétique. Mes travaux de thèse permettent donc de caractériser la recombinaison méiotique en Lacaune et ont pour objectif de contribuer à la compréhension de la recombinaison génétique chez le mouton : création de cartes génétiques, localisation de la recombinaison et des crossing-overs, observation de la variation du taux de recombinaison et de son déterminisme génétique.

Ce manuscrit est composé de quatre grandes parties. La première partie est une étude bibliographique présentant la recombinaison génétique, les phénotypes étudiés et leur déterminisme génétique. La seconde partie présente la création de cartes génétiques haute résolution, la troisième partie porte sur l'étude de la variation du taux de recombinaison et de son déterminisme génétique. Finalement, la dernière partie est une application pratique de l'utilisation des cartes génétiques. Elles peuvent être utilisées pour créer des puces basse densité qui pourront être utiles en sélection génomique afin de permettre une meilleure imputation à partir des puces avec une densité moyenne de marqueurs.

Synthèse Bibliographique

Chapitre 1 : La recombinaison méiotique ; biologie fonctionnelle, phénotypes étudiés et déterminisme génétique associé

I. Biologie fonctionnelle de la recombinaison

I. 1. *La recombinaison méiotique : définition et rôle*

Historiquement, la recombinaison génétique définit tous phénomènes conduisant à la formation, dans une cellule fille, de fragments issus du clivage de chromosomes qui sont ensuite reliés pour donner de nouvelles combinaisons, différentes de celles observées dans les cellules parentales (Lodish *et al.*, 2000). Il existe plusieurs manières de définir la recombinaison génétique. Cela peut faire référence au mécanisme moléculaire, à un processus contribuant à l'évolution des espèces et à la diversité génétique, ou bien à un processus de réparation de l'*ADN*[1] (Acide Désoxyribonucléique). En biologie moléculaire, elle correspond à la liaison, puis à l'échange entre deux molécules d'*ADN* de régions différentes, plus ou moins éloignées. Ce phénomène est présent chez tous les organismes cellulaires et chez quelques virus (Lodish *et al.*, 2000).

La recombinaison est donc un terme très général, qu'on peut préciser en fonction de son type. La recombinaison méiotique dite « homologue » correspond à l'échange de séquences nucléotidiques identiques, qui peuvent être situées sur n'importe quel chromosome.

La réparation des cassures d'*ADN*, *DSBs* (Double-Strand Breaks), joue un rôle clé dans la recombinaison génétique, car c'est elle qui l'initie. Ces *DSBs* peuvent être volontairement induites par une cellule (Lodish *et al.*, 2000). L'un des rôles essentiels de la recombinaison est le maintien de la diversité génétique pour les espèces sexuées. En effet, en permettant une ségrégation correcte des chromosomes au cours de la méiose, elle conduit à la formation de gamètes exempts

1 Les abréviations en italique renvoient au lexique des abréviations.

d'anomalies génétiques et contribue au brassage génétique (Cromie et Smith, 2007).

I. 2. *La recombinaison génétique a lieu au cours de la méiose*

La méiose est un processus universel qui permet la formation des gamètes pour les organismes eucaryotes, elle est donc essentielle à la reproduction sexuée des espèces. Elle affecte alors les cellules de la lignée germinale et permet le passage de l'état diploïde à l'état haploïde conduisant à la formation des spermatozoïdes et des ovocytes. La méiose est constituée de deux divisions consécutives sans réplication d'*ADN* entre elles. La première est dite réductionnelle car une cellule au préalable diploïde (gonie, à 2n chromosomes) va se diviser pour donner deux cellules filles haploïdes. Cela permet de faire en sorte que le nombre de chromosomes d'une espèce reste le même d'une génération à l'autre. La deuxième division, équationnelle, ressemble à la mitose et va conduire à la formation de quatre cellules filles (gamètes, à n chromosomes), contenant chacune une seule chromatide issue des chromosomes de départ. En revanche, chez les Mammifères femelles, il n'y a production que d'un seul gamète, les trois autres cellules étant expulsées sous la forme de globules polaires au cours de l'ovogenèse.

La méiose a été étudiée pour la première fois en 1766, chez le tabac, par Kölreuter, avant que ne soit découverte la structure de la molécule d'*ADN*. Son rôle fondamental dans la reproduction sexuée a été découvert une centaine d'années plus tard par Weismann (1890).

Préalablement à toute division cellulaire, notamment la méiose, l'*ADN* est répliqué (réplication pré-méiotique), afin de devenir bichromatidien. Les chromosomes sont également arrangés par paires d'homologues. A la suite de cela, la cellule entre en méiose. Chacune des deux divisions est constituée de quatre phases : prophase, métaphase, anaphase et télophase (**voir Figure 1**).

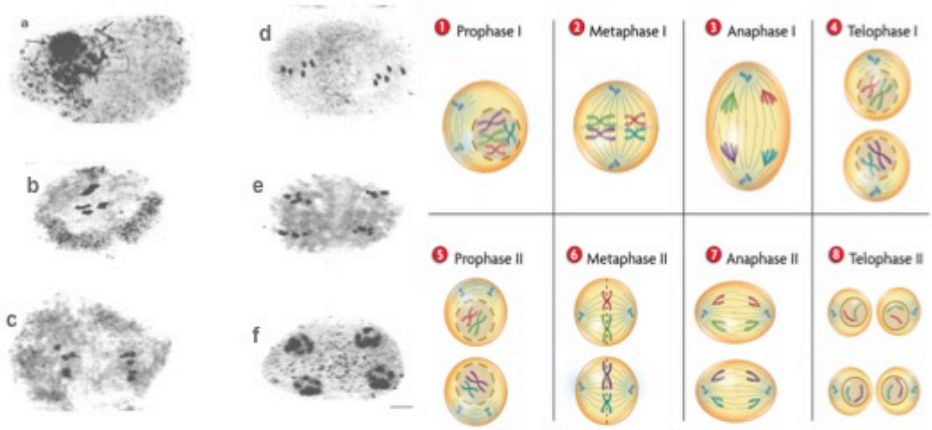

Figure 1 : Différentes phases de la méiose en photographie (gauche) et en schéma (droite), d'après Science Notebook : http://tmsseventhgradeteam.weebly.com/notebook/archives/03-2016).

Les différentes phases de la méiose marquée au *DAPI* (DNA fluorochrome 4',6-diamidino-2-phenylindol) combiné au *HIA* (Haematoxylin-Iron Alum) chez *Arabidopsis thaliana* (d'après Ross *et al.*, 1996). **a-b-c**/ Première division de méiose. **a**/ Prophase I : les chromosomes sont visibles sous la forme de longs filaments en cours de condensation. Les flèches indiquent les nucléoles. **b**/ Métaphase I : 5 bivalents très condensés en train de s'aligner sur la plaque métaphasique à l'aide des kinétochores. Chaque bivalent est lié par les chiasmata issus des échanges de chromatides (crossing-overs) qui ont eu lieu à la phase précédente. **c**/ Anaphase I : séparation des bivalents et migration de chaque chromosome vers un pôle de la cellule. Les chromatides sœurs sont toujours reliées grâce aux crossing-overs. **d-e-f**/ Deuxième division de méiose. **d**/ Métaphase II : deux groupes de 5 chromosomes alignés sur la plaque équatoriale. **e**/ Anaphase II : séparation des 5 chromatides et migration de chacune vers un pôle de la cellule. **f**/ Tétrade de 4 nucléi haploïdes avant formation des 4 cellules filles suite à la Télophase II.

I. 3. *Détails de la Prophase I : étape essentielle de la recombinaison méiotique*

C'est la phase la plus longue et la plus complexe de la méiose, elle représente 90 % de la durée totale du mécanisme. Au cours de cette étape, il y a condensation progressive des chromosomes qui s'apparient ensuite par paire pour former des bivalents grâce au complexe synaptonémal (*CS*). Il s'agit d'un complexe protéique permettant l'association des chromosomes homologues entre eux. Cet appariement est un synapsis (**voir Figure 2**). Le *CS* est une structure

tripartite comprenant des éléments axiaux, composés par les protéines *SYCP1*, *SYCP2* et *SYCP3* (Capilla *et al.*, 2016). La création et la réparation des *DSBs* a lieu au niveau de cette structure chromosomique axiale. C'est pendant la Prophase I qu'ont lieu les non-crossing-overs (*NCOs*) et les crossing-overs (*COs*), ces derniers permettant le brassage intrachromosomique. Les crossing-overs ne modifient pas la succession des gènes, mais uniquement la répartition des allèles[2].

Plusieurs évènements ont lieu au cours de la Prophase I, notamment la condensation des chromosomes, la formation des bivalents et la recombinaison homologue. Etant donné son importance et grâce à des techniques de microscopie, il a été possible de la découper en cinq étapes : Leptotène, Zygotène, Pachytène, Diplotène et Diacinèse (**voir Figure 2**).

Etape Leptotène

A ce stade, les chromosomes sont dupliqués, mais encore en cours de condensation. Ils sont donc encore longs et fins, d'où le nom donné à cette étape (du grec « *leptos »* pour « fin » et « *tenôn* » pour « tendre », « étirer »). C'est à ce moment que se mettent en place les premiers nodules de recombinaison (*EN*, pour « Early Nodules »), prémices de la recombinaison homologue, avec formation des *DSBs* et début de leur réparation. Les nodules de recombinaison sont des structures protéiques au sein desquelles a lieu la recombinaison homologue, elles sont étroitement associées au complexe synaptonémal. Dans la plupart des espèces, les extrémités des chromosomes sont attachées à l'enveloppe nucléaire via la plaque d'attachement, cette configuration, appelée « bouquet télomérique» semble aider à l'alignement futur des chromosomes et permet la transition du stade Leptotène au stade Zygotène (Baudat *et al.*, 2013).

Etape Zygotène

Le stade Zygotène, du grec « *zugon* », signifiant « paire », est caractérisé par le début de l'appariement des chromosomes homologues en vue de former les bivalents. Les chromosomes sont beaucoup plus condensés, donc plus épais et plus courts. L'appariement des chromosomes

2 Un allèle, du grec *allelos* « l'un, l'autre », désigne chacune des différentes formes ou versions possibles d'un même gène.

est permis par la formation du *CS*. Il se met en place au début au niveau de l'enveloppe nucléaire, puis progresse le long des chromosomes à la manière d'une « fermeture éclair ». Une mutation au niveau des protéines responsables de sa formation peut conduire à une impossibilité pour les chromosomes homologues de s'apparier (Shin *et al.*, 2010). Le rôle exact de ce complexe n'est pas encore totalement connu, mais il pourrait être important au niveau de l'interférence des crossing-overs. L'interférence se définit comme étant la probabilité que lorsqu'un crossing-over apparaît dans une région du génome, la présence d'un deuxième crossing-over dans la même région est réduite. En général, elle est complète pour des régions très proches physiquement, puis elle diminue avec la distance au site d'initiation du premier crossing-over (Jones et Franklin, 2006). Ce sont Sturtevant et Muller qui ont les premiers définis l'interférence, respectivement en 1915 et 1916. Au cours de cette étape, la résolution des *DSBs* se poursuit avec la formation d'intermédiaires de recombinaison et de boucles qui seront ensuite résolues, ou non, en crossing-overs (Baudat *et al.*, 2013).

Etape Pachytène

Durant cette étape, la condensation des chromosomes se termine, ils apparaissent alors assez épais, d'où le nom de ce stade (du grec « *pakhus* », signifiant « gros »). Le complexe synaptonémal est terminé et l'appariement des chromosomes homologues est alors complet. C'est souvent l'étape la plus longue de la Prophase I ; 8 jours chez la souris et 15 chez l'Homme (Adler, 1996). Au cours de ce stade, il y a formation de nodules tardifs de recombinaison (*LN*, pour « Late Nodule »), qui pourraient correspondre aux sites des crossing-overs, notamment car le nombre de *LNs* correspond au nombre d'échanges réciproques entre les chromosomes homologues, de plus ils sont distribués de manière non aléatoire et se situent au niveau des sites des échanges (Carpenter, 1979). Les *LNs* sont notamment associés à un complexe protéique nécessaire à la formation des crossing-overs (Moens *et al.*, 2001). Les produits finaux de la recombinaison (crossing-overs et non-crossing-overs) sont générés à la fin de cette étape (Baudat *et al.*, 2013).

Etapes Diplotène et Diacinèse

Le stade Diplotène, du grec « *diploos* », pour « double », est marqué par le début de la

séparation des chromosomes homologues, qui est due à la dissolution du complexe synaptonémal. Les bivalents peuvent donc s'éloigner d'une certaine distance, mais ils restent reliés au niveau des chiasmata, stabilisés par cohésion entre les chromatides sœurs.

Le stade « Diacinèse » (« *dia* » pour « séparation » et « *kinêsis* » pour « mouvement » en grec) traduit la fin de la Prophase I et la transition avec la Métaphase I. Les chromosomes ont atteint leur maximum de compaction et se détachent de l'enveloppe nucléaire. Les chromatides sœurs de chaque paire sont reliées par leur centromère, tandis que les chromatides non-sœurs comportant un crossing-over sont reliées par leur chiasma.

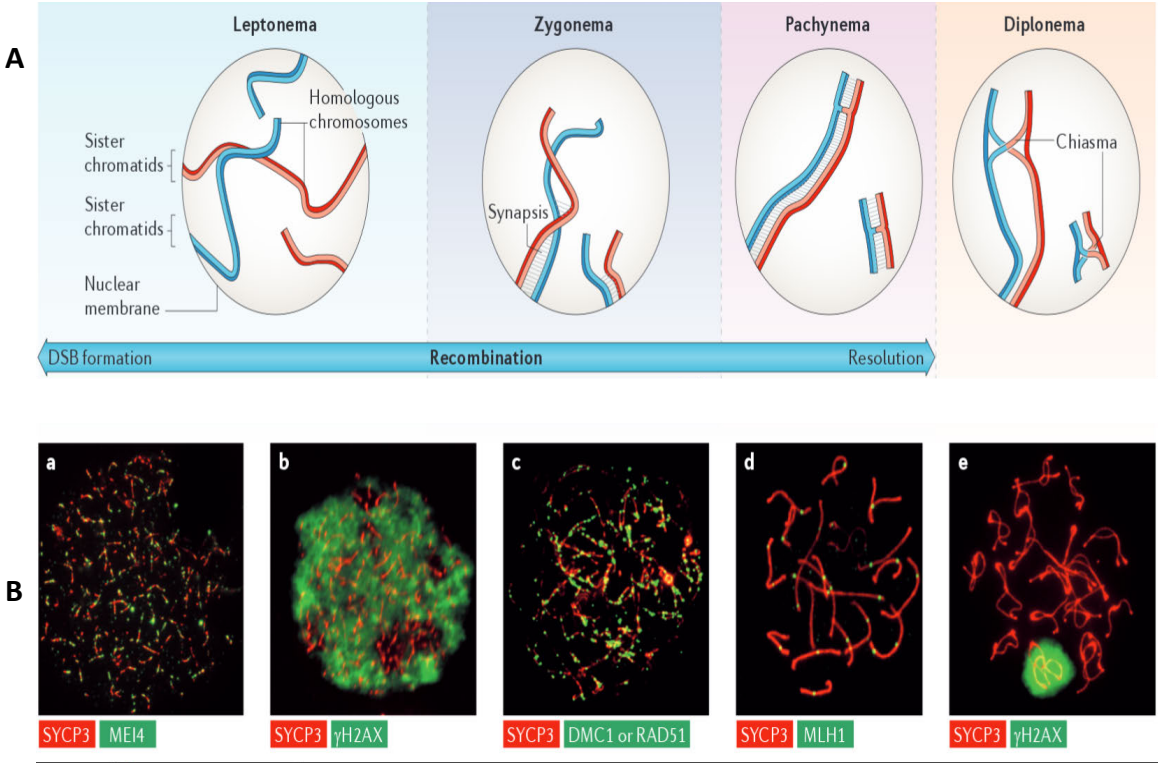

Figure 2 : Les différentes étapes de la Prophase I (D'après Baudat et al., 2013)

A/ Schématisation des différentes phases de la Prophase I. Au stade Leptotène, les chromosomes sont en cours de condensation. Les chromatides sont très étroitement accolées, les chromosomes semblent donc avoir un unique brin. Les *DSBs* se forment à cette étape. Au stade Zygotène, la formation du complexe synaptonémal commence à associer fortement les chromosomes homologues et les *DSBs* commencent à être résolues. L'association du *CS* est complète au stade Pachytène et des crossing-overs peuvent apparaître suite à la formation des nodules de recombinaison à l'issue de ce stade. Le *CS* est dissous au stade Diplotène, les chromosomes homologues sont reliés entre eux au niveau des chiasmata, illustrant les sites de formation des crossing-overs. Enfin, au stade Diacinèse, les chromosomes ont atteint leur condensation maximale. **B**/ Observations cytologiques des différents stades de la Prophase I avec immunolocalisation de différentes protéines impliquées dans la résolution des *DSBs* (en vert) ou dans la formation du *CS* (en rouge).

I. 4. *La formation des crossing-overs*

Les évènements de recombinaison ; crossing-overs, et conversion génique, ainsi que les non-crossing-overs, sont issus de la recombinaison méiotique suite à la formation des *DSBs.* C'est donc grâce à ces lésions et à leur réparation qu'a lieu la recombinaison homologue.

Dès 1909, Janssens a proposé un modèle pour comprendre la formation des crossing-overs et des non-crossing-overs : la rupture, puis la réparation par jonction des chromosomes homologues (Janssens, 1909). Cela fut démontré plusieurs années après par Game et collaborateurs (Game *et al.*, 1989).

I. 4. a. *Initiation des cassures double brin*

La formation des cassures double brin est catalysée par l'enzyme *SPO11* au cours du stade Leptotène de la Prophase I de méiose (**voir Figure 3**). C'est une protéine très conservée dont le rôle a pu être démontré chez la souris, où de forts niveaux de phosphorylation *H2AX*, *SPO11*-dépendants, ont été détectés à ce stade. De plus, des molécules oligo-*SPO11* libres, résultant du clivage et de la résection des extrémités 5', sont détectées dans les testicules de souris non mutées, ce qui étaie l'hypothèse de la présence de *SPO11* au cours de la méiose (Baudat et de Massy, 2007). Son importance a pu être démontrée grâce à l'étude de mutants chez la levure. En effet, lorsque *SPO11* est absente, les levures ne présentent pas de *DSBs*, pas ou peu d'événements de recombinaison et une perturbation, voire une absence complète de formation du *CS* (Keeney, 2001). Cette protéine serait recrutée par le gène PR domain-containing 9 (*PRDM9*). Il code pour une protéine ayant un rôle d'histone méthyltransférase (Baudat *et al.*, 2013) et contenant, entre autres, un domaine spécifique avec plusieurs doigts de zinc. Ces doigts de zinc permettraient à la protéine *PRDM9* de se fixer sur l'*ADN*, au niveau de zones particulières du génome. La fixation de *PRDM9* au niveau de ces régions induirait donc la formation des *DSBs* par le biais du recrutement de *SPO11.* Cependant, l'absence de *PRDM9* chez des souris mutantes, n'empêche pas la formation des *DSBs*. En revanche, elles se forment à des endroits totalement différents de ceux des souris sauvages (Baudat *et al.*, 2013) ; se regroupant majoritairement au niveau des sites de démarrage de la transcription ou des promoteurs. Néanmoins, bien que les *DSBs* se forment chez les souris mutantes, leurs spermatocytes ne se forment pas correctement au cours de la méiose et les animaux sont infertiles. *PRDM9* semble donc avoir deux fonctions principales : spécifier les sites de formation des *DSBs* et assurer leur réparation (Baudat *et al.*, 2013).

Plusieurs protéines sont nécessaires à la réparation des *DSBs*, ainsi les oligos-*SPO11* semblent résulter du clivage du complexe *ADN-SPO11* par l'action d'une endonucélase ce qui génère deux extrémités 5'. Bien que cette endonucléase ne soit pas encore totalement connue,

des études récentes ont montré qu'il y avait formation d'un complexe avec la protéine topoisomérase *TOPOVIBL* (Vrielynck *et al.*, 2016). Cette protéine conduit au clivage et à la ligation de brins d'*ADN* chez les bactéries archées (Robert *et al.*, 2016), et des orthologues ont pu être mis en évidence chez les plantes et les Mammifères. *TOPOVIBL* est nécessaire à la formation des *DSBs* car des souris mutantes pour cette protéine ont des testicules plus petits que les souris non mutantes. Il semble donc que le complexe *SPO11-TOPOVIBL* soit nécessaire à la formation des *DSBs*, peut-être en permettant cette étape de clivage (Robert *et al.*, 2016). *SPO11* est une molécule clé pour l'initiation de la réparation des *DSBs*, car il a été montré que, chez des souris mutantes pour l'enzyme, il n'y avait pas de méiose (Romanienko et Camerini-Otero, 2000). Les extrémités 5' ainsi créées sont ensuite réséquées pour conduire à de longs filaments simple brin avec des extrémités 3'-OH qui pourraient ensuite envahir la deuxième chromatide homologue, et ce grâce au complexe *RAD50-MRE11-XRS2* (Smith et Nicolas, 1998). Cet envahissement conduit à la formation de complexes bimoléculaires, comprenant une boucle D-loop et une jonction de Holliday (Smith et Nicolas, 1998). Plusieurs gènes impliqués dans l'échange de brins participent à la formation de ces complexes, notamment *RAD51*, *RAD55*, *RAD57* et *DMC1*. Chez les plantes, des mutants pour ces gènes produisent généralement des spores non viables, d'où le rôle clé de ces molécules (Smith et Nicolas, 1998). *DMC1*, contrairement aux autres protéines, est spécifique de la recombinaison méiotique. Elle va aider à la formation d'interactions entre les chromosomes homologues, il s'agit du « biais inter-homologue », lors de la création de la boucle en détectant des séquences complémentaires. Une fois que l'échange des brins est terminé, les protéines *RAD51* et *DMC1* sont éliminées. La boucle va pouvoir migrer et de l'*ADN* sera synthétisé à partir de l'extrémité 3' qui sert d'amorce. C'est à partir de cette étape qu'il y a divergence et différentes possibilités de réparation des cassures, conduisant à la formation de crossing-overs de classe I, de crossing-overs de classe II ou de non-crossing-overs.

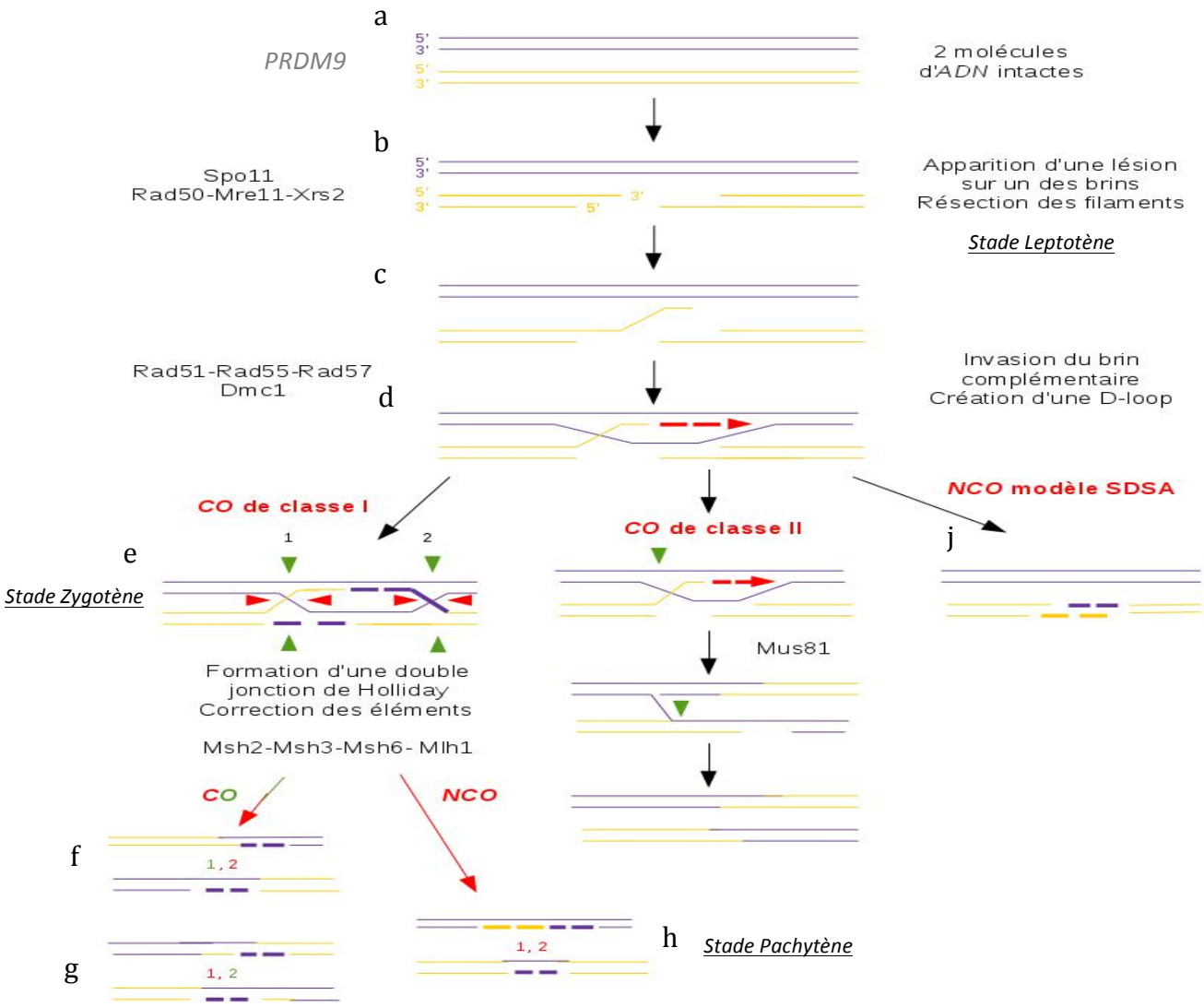

Figure 3 : Mécanisme de formation des différents évènements de recombinaison (d'après Saintenac, 2012)

a/ Deux molécules d'*ADN* homologues intactes, une première en violet, la deuxième en jaune. **b**/ Apparition d'une lésion sur le chromosome jaune suite à la fixation de *PRDM9*, le chromosome homologue reste intact ; formation d'une cassure double-brin qui initie la recombinaison méiotique grâce à l'enzyme *SPO11*. Résection des extrémités 5' pour donner des extrémités 3'-OH par l'action du complexe *RAD50-MRE11-XRS2*. **c-d**/ Invasion du brin complémentaire par l'extrémité 3' libérée et formation d'une D-loop qui s'étend grâce aux enzymes *RAD51* et/ou *DMC1* afin de permettre la synthèse d'*ADN*. A partir de là, plusieurs possibilités de résolution de la cassure. **e**/ Modèle classique : invasion de la D-loop par la deuxième extrémité 3' et liaison avec l'ancien brin envahissant, ce qui aboutit à la formation d'une double jonction de Holliday. Stabilisation de la structure et correction des erreurs de complémentarité grâce à

l'action du complexe *MSH4-MSH5-MLH1*. Selon la résolution des jonctions (coupures schématisées par les flèches vertes et rouges), le modèle classique peut conduire à la formation de crossing-overs de classe I ; coupure de la jonction 1 en fonction des flèches vertes et coupure de la jonction 2 en fonction des flèches rouges (/**f**) ou coupure de la jonction 1 en fonction des flèches rouges et coupure de la jonction 2 en fonction des flèches vertes (/**g**). Au contraire, si la double jonction s'annule, on obtient des non-crossing-overs (/**h**). Une autre voie de résolution des cassures double-brin conduit à la formation de crossing-overs de classe II, non soumis à l'interférence, sous l'action de l'enzyme *MUS81*, avec formation uniquement d'une simple jonction de Holliday (/**i**). **j**/ Enfin, la résolution des cassures double brin peut conduire uniquement à la formation de non-crossing-overs avec le modèle *SDSA* (Synthesis-Dependent Strand-Annealing). Le brin envahissant se retire de la D-loop, ce qui l'élimine, et s'apparie avec son brin complémentaire. La synthèse d'*ADN* permet de combler les cassures.

I. 4. b. *La résolution des cassures double brin*

I. 4. b. a. La formation de crossing-overs de classe I

Dans ce premier cas, le brin 5' de la deuxième extrémité de la cassure envahit la boucle et se lie avec le précédent brin envahissant, ce qui conduit à la formation d'une double jonction de Holliday. La résolution de cette boucle peut conduire à la formation de crossing-overs de classe I ou à des non-crossing-overs (Baudat et de Massy, 2007). Il a été démontré chez la souris qu'il y a un large excès du nombre de foci, c'est-à-dire des zones ou des sites (sorte de « points ») où s'accumulent des protéines spécifiques marquées par immunofluorescence, *RAD51/DMC1*, ce qui suggère que les non-crossing-overs sont majoritaires par rapport aux crossing-overs chez les Mammifères (Baudat et de Massy, 2007). Le nombre de crossing-overs peut cependant être sous-estimé, car leur détection dépend de la présence de marqueurs polymorphiques. Chez la levure *S. cerevisiae*, plusieurs gènes sont nécessaires à la formation des crossing-overs de classe I : *ZIP1*, *ZIP2*, *ZIP3* (orthologue chez la souris : *rnf212*), *ZIP4* (orthologue chez la souris : *Tex11*), *SPO16*, *MSH4*, *MSH5* et *MER3* (orthologue chez la souris : *Hfm1*) (appelés collectivement « *ZMM* »). Des individus mutants semblent avoir une réduction du nombre de crossing-overs, alors que les non-crossing-overs ne sont pas affectés (Baudat et de Massy, 2007, Baudat *et al.*, 2013). La formation de crossing-overs nécessite également la présence de *MLH1* et *MLH3*, les mutants présentant également une réduction du nombre de crossing-overs, mais moindre (Wang *et al.*, 1999). *In vivo*,

l'hétérodimère *MSH4-MSH5* pourrait permettre la stabilisation de la D-loop et de la double jonction de Holliday et participer à la maturation et/ou à la résolution de la double jonction (Hoffmann et Borts, 2004). *MLH1* pourrait jouer un rôle dans la résolution et la correction des éléments formés au cours de la recombinaison homologue, cependant, il semble que le complexe *MLH1-MLH3* ne soit pas nécessaire pour la formation des non-crossing-overs chez les Mammifères, contrairement à la formation des crossing-overs (Baudat et de Massy, 2007). De plus, chez les Mammifères, il semble que *MSH4* ne soit pas spécifique de la formation des crossing-overs, mais influe également sur les non-crossing-overs. Une hypothèse de niveau d'interférence de certaines protéines, notamment le complexe *MLH1-MSH4*, pourrait expliquer la résolution vers un crossing-over ou un non-crossing-over (Baudat et de Massy, 2007).

I. 4. b. b. La formation des crossing-overs de classe II

Les crossing-overs de classe II se forment à partir d'une voie différente de celle des crossing-overs de classe I, présentée précédemment, où il n'y a pas formation d'une double jonction de Holliday. Elle fait appel à d'autres protéines, notamment le complexe *MUS81-MMS4* chez *S. cerevisia (MUS81-EME1* chez *S. pombe*) (Hollingsworth et Brill, 2004), *MUS81* étant une endonucléase. Ces protéines ont été proposées car il a été montré, que même partiellement purifiées, elles étaient capables de cliver une jonction de Holliday intacte chez l'Homme et *S. pombe*. De plus, chez des individus mutants, le nombre de crossing-overs est diminué, alors qu'il est à peine affecté chez les non-crossing-overs (Hollingsworth et Brill, 2004). Le mécanisme d'action exact de *MUS81* est encore soumis à controverse. L'enzyme pourrait avoir un rôle au début de la formation des *DSBs*, lors de l'invasion du brin homologue, ou plus tard, au moment de la résolution des jonctions de Holliday (Holloway *et al.*, 2008). Chez *S. pombe*, il semble que *MUS81* soit capable de cliver des jonctions intactes de Holliday, mais ce substrat ne semble pas être le même chez *S cerevisiae* et les eucaryotes évolués (Holloway *et al.*, 2008). Un premier modèle a été proposé par Li et Brill (2005) où *MUS81* clive l'extrémité 3' du brin intact, ce qui peut conduire à une double jonction de Holliday sous le contrôle de *RAD51*. Cette structure serait résolue en crossing-over grâce à la synthèse d'*ADN* et à une ligation (Li et Brill, 2005). Dans un deuxième modèle, proposé par De Los Santos et collaborateurs, *MUS81* cliverait le brin déplacé au cours de l'invasion pour conduire à une D-loop et à une jonction simple de Holliday (De Los Santos

et al., 2003).

Les deux voies de formation des crossing-overs diffèrent selon le type de structure créé : jonction simple de Holliday ou jonction double, selon les protéines impliquées ou encore selon leur réponse au phénomène d'interférence. En effet, les crossing-overs de classe II semblent se former indépendamment les uns des autres, alors que les crossing-overs de classe I sont influencés par la présence ou non d'autres crossing-overs et sont donc souvent plus espacés (Holloway *et al.*, 2008). Chez la plupart des Eucaryotes, les crossing-overs sont très majoritairement de classe I : environ 90 % chez la souris (Guillon *et al.*, 2005), 85 % chez *A. thaliana* (Franklin *et al.*, 2006) ou encore 70 % chez *S. cerevisiae* (Argueso *et al.*, 2004). En revanche, chez *S. pombe*, quasiment tous les crossing-overs sont de classe II (Holloway *et al.*, 2008).

I. 4. b. c. La formation de Non-Crossing-overs

On fait souvent référence aux non-crossing-overs en tant que « conversion génique », or les conversions géniques correspondent plutôt à des sites où les évènements de recombinaison résultent en une ségrégation non-mendélienne d'un ou plusieurs marqueurs génétiques (Bishop et Zickler, 2004). Les non-crossing-overs sont majoritaires par rapport aux crossing-overs, en effet, ces derniers ne représentent que 11 % des évènements de recombinaison inter-homologues (Rosu *et al.*, 2011). Les non-crossing-overs résultent d'un mécanisme appelé « Synthesis-Dependent Strand Annealing » (*SDSA*) où le brin 3' envahissant s'étend sous l'action d'une polymérase. Puis, il retourne s'hybrider avec son brin complémentaire d'origine (Baudat et de Massy, 2007). Ceci peut être dû à l'action de topoisomérases ou d'hélicases qui démantèlent la boucle (Pâques et Haber, 1999). La continuité de chaque brin est assurée par la synthèse d'*ADN*. Ce modèle *SDSA* a été suggéré chez la drosophile par Nassif *et al.* en 1994 (Nassif *et al.*, 1994). Il existe des mécanismes similaires dans d'autres espèces, telles que les Mammifères ou encore *E. coli* (Pâques et Haber, 1999). Il semblerait que les hétéroduplex formés au cours de ce mécanisme apparaissent au même moment que les doubles jonctions de Holliday de la voie des crossing-overs de classe I (Youds et Boulton, 2011). Chez *C. elegans*, une protéine a été découverte ayant un rôle dans le modèle *SDSA*. Il s'agit de l'anti-recombinase *RTEL-1* (Youds et Boulton, 2011). Les mutants pour cette enzyme montrent de multiples crossing-overs par chromosome, alors que les individus

sauvages ne présentent qu'un seul crossing-over par chromosome. *In vitro*, la protéine a la capacité de dissocier la D-loop ; l'hypothèse serait donc qu'elle agisse de même dans le modèle *SDSA*, permettant ainsi au brin envahissant de se réassocier facilement avec son complémentaire d'origine. Actuellement, il s'agit du seul exemple de protéines « anti-crossing-overs », bien qu'il semblerait qu'il existe chez *S. cerevisiae*, une enzyme, *MPH1* (Mutator Phenotype 1) qui aurait des similarités biochimiques avec *RTEL-1* (Youds et Boulton, 2011).

Il est souvent déclaré qu'il y a une majorité de non-crossing-overs par rapport aux crossing-overs, or chez *S. cerevisiae*, l'étude d'une carte génétique haute résolution, a permis de détecter 90,5 crossing-overs contre seulement 46,2 non-crossing-overs par méiose (Mancera *et al.*, 2008). Cependant, sur ces crossing-overs observés, 30,1 % d'entre eux ont lieu entre deux marqueurs consécutifs, ce qui ne permet pas de détecter certains non-crossing-overs. L'équipe a donc évalué cette fraction non détectée et arrive ainsi à 66,1 non-crossing-overs par méiose, ce qui correspond à environ 140 à 170 *DSBs* par méiose (Mancera *et al.*, 2008). Ce faible nombre de non-crossing-overs est sûrement dû à une sous-estimation causée par la difficulté de leur détection. Dans d'autres espèces, en revanche, telles que la souris, l'Homme ou *A. thaliana*, un excès de non-crossing-overs est observé et est caractérisé de manière cytologique : utilisation d'anticorps contre les protéines *Rad51* et/ou *Dmc1*. Ces études cytologiques ont permis de montrer qu'il y avait 10 à 40 fois plus de sites de réparation des *DSBs* (identifiés par les protéines), que de crossing-overs (Baudat et de Massy, 2007). Cela pourrait donc suggérer que les *DSBs* non résolues en crossing-overs, le sont en non-crossing-overs, d'autant plus que le nombre total de *DSBs* correspond à la somme des crossing-overs et des non-crossing-overs (Mancera *et al.*, 2008).

Tableau 1 : Récapitulatif des protéines intervenant dans le processus de la recombinaison

Protéine	Stade d'action	Rôle(s)
PRDM9	Début du stade Leptotène de la Prophase I.	Recrutement de *SPO11* et induction des *DSBs*.
SPO11	Leptotène de la Prophase I.	Formation des *DSBs*.
TOPOIVB	Leptotène de la Prophase I.	Formation des *DSBs*.
Complexe *RAD50-MRE11-XRS2*	Leptotène de la Prophase I.	Résection des extrémités des brins d'*ADN*.
Complexe *RAD51, RAD55, RAD57 et DMC1*	Zygotène de la Prophase I.	Invasion du brin d'*ADN* réséqué et formation de D-loop.
Complexe *ZIP1, ZIP2, ZIP3* (*RNF212*)	Pachytène de la Prophase I.	Formation de la double jonction de Holliday.
Complexe *MSH4-MSH5*	Pachytène de la Prophase I.	Stabilisation de la D-loop et de la double jonction de Holliday, maturation et/ou résolution de la double jonction.
Complexe *MLH1-MLH3*	Pachytène de la Prophase I.	Résolution et correction des évènements de recombinaison, nécessaires à la formation des crossing-overs de type I.
Complexe *MUS81-MMS4*	Pachytène de la Prophase I.	Clivage de la double jonction de Holliday menant à une jonction simple de Holliday. Nécessaire à la formation de crossing-overs de type II.
RTEL-1/MPH1	Pachytène de la Prophase I.	Protéines « anti-crossing-overs » conduisant à la formation de non-crossing-overs grâce au modèle de résolution *SDSA*.

I. 4. c. *Conclusion intermédiaire*

La Prophase I de la méiose est donc une des étapes les plus importantes puisque c'est à ce moment que se forment les crossing-overs et les non-crossing-overs. Ces derniers, à l'inverse des crossing-overs, ne sont pas considérés comme des évènements de recombinaison car ils n'affectent pas la ségrégation des marqueurs (sauf s'ils sont à l'origine de conversion génique). Les

crossing-overs permettent une bonne ségrégation des chromosomes et la formation de gamètes exempts de problèmes génétiques. Ces crossing-overs peuvent être de deux types selon qu'il y ait ou non création d'une double jonction de Holliday. La formation des crossing-overs suit la résolution des cassures double-brin d'*ADN* selon un mécanisme relativement complexe qui fait intervenir de nombreuses molécules (**voir Tableau 1**). Cependant, les crossing-overs ne sont pas les évènements de recombinaison majoritaires, en réalité, ils ne représentent que 11 % des évènements totaux. Les *DSBs* seraient, en fait, plus particulièrement résolues en non-crossing-overs.

II. Etude de la recombinaison méiotique

II. 1. *Méthodes d'estimation des cartes de recombinaison*

II. 1. a. *Différentes méthodes utilisées pour étudier la recombinaison*

La recombinaison peut être estimée grâce au nombre d'évènements de recombinaison (crossing-over) sur l'ensemble du génome, mais on peut aussi étudier la distribution de ces évènements de recombinaison, c'est-à-dire le rapport entre le taux d'évènements de recombinaison et une distance physique.

Plusieurs techniques sont utilisables pour l'étude de la recombinaison. Il existe notamment des méthodes de cytogénétique qui peuvent permettre d'estimer les distances physiques. Ainsi que des méthodes plus indirectes et quantitatives avec l'utilisation de données de génotypages de populations d'individus. Grâce à l'utilisation de puces à marqueurs, il est possible d'estimer des distances génétiques, pour des distances physiques connues. Il est ensuite possible de comparer ces deux types de distance.

II. 1. a. a. Approches cytogénétiques et moléculaires

Des techniques de cytogénétique ont permis de repérer « physiquement » et directement les évènements de recombinaison sur les chromosomes. Connaître la distribution des crossing-overs le long des chromosomes est important, car cela permettrait de mieux comprendre leur

régulation (Anderson *et al.*, 1999). Les premières études cytologiques ont permis de visualiser grossièrement la distribution des crossing-overs, notamment grâce à l'étude des chiasmata sur les chromosomes bivalents. Cela a été réalisé au stade Diplotène, en particulier chez le criquet et la sauterelle, et a notamment permis de montrer que les crossing-overs ne sont pas distribués aléatoirement sur les chromosomes (Anderson *et al.*, 1999). A la suite de cela, des techniques de microscopie électronique ont permis d'étudier les nodules tardifs de recombinaison, associés à la formation des crossing-overs et présents le long du complexe synaptonémal lors du stade Pachytène (Anderson *et al.*, 1999). Ces nodules sont plus rarement observés chez les Mammifères, donc il y a peu de cartes physiques qui ont été réalisées pour ces animaux à l'aide de cette technique.

Pour les Mammifères, une possible alternative est la détection de protéines spécifiques à l'aide de techniques d'immunologie fluorescente. Il s'agit de protéines qui interviennent dans le mécanisme de recombinaison méiotique et qui sont détectées à l'aide d'anticorps spécifiques. La plus utilisée, notamment chez la souris ou la levure, est la protéine *MLH1* (MutL Homolog) qui se propage sur le complexe synaptonémal (Anderson *et al.*, 1999). Il s'agit d'une protéine de réparation de l'*ADN*. Il a été démontré par Baker *et al.*, (1996) une bonne corrélation entre la localisation de cette protéine au stade Pachytène et la présence de crossing-overs. Ces différentes techniques ont permis de montrer que la localisation des crossing-overs le long des bivalents n'est pas homogène. Notamment, certaines régions, comme les télomères, les zones péricentriques ou constituées de séquences répétées, sont quasiment dépourvues de crossing-overs. Elles permettent également d'estimer le nombre de crossing-overs sur les chromosomes.

Il est également possible de rechercher d'autres protéines impliquées dans la réparation des *DSBs*, telles que *DMC1* ou *SPO11*. Pour la détection de *DMC1*, la chromatine est extraite de testicules adultes puis immuno-précipitée avec des anticorps spécifiques de cette protéine et associés à un simple brin d'*ADN* (Pratto *et al.*, 2014). Les fragments d'*ADN* immuno-précipités sont ensuite analysés par séquençage haut-débit, ce qui permet de localiser des zones de forte recombinaison avec une résolution de l'ordre d'1 Kb. Avec cette méthode, il est normalement possible de détecter tous les crossing-overs, ainsi que les non-crossing-overs (Baudat *et al.*, 2013). Chez la levure, le séquençage des oligonucléotides de *SPO11*, libérés au cours de la formation des *DSBs*, permet de détecter les *DSBs* dépendantes de *SPO11*, avec une résolution de l'ordre du

nucléotide (Baudat *et al.*, 2013). L'étude de protéines impliquées dans la formation du complexe synaptonémal : *SYCP1*, *SYCP2* ou *SYCP3* (Capilla *et al.*, 2016) permet, par ailleurs, d'en connaître la longueur.

Ces techniques, dites de « *ChIP*-sequencing » (chip-seq) (Chromatin Immuno-Precipitation) sont toujours largement utilisées afin de comprendre les processus de la méiose et les protéines qui y jouent un rôle. En revanche, elles présentent quelques limites, notamment la nécessité de disposer de tissus testiculaires ou fœtaux (Capilla *et al.*, 2016). De plus, ce sont des techniques assez lourdes, coûteuses et parfois peu précises car de l'ordre de la mégabase, ce qui ne permet pas de localiser précisément les crossing-overs.

II. 1. a. b. Approches indirectes et statistiques

L'étude de la génétique a commencé avec Mendel dans les années 1860, mais ces travaux sont restés assez inaperçus à l'époque. Il faudra attendre les années 1900 pour qu'ils soient redécouverts et réétudiés. Ainsi, l'allemand Carl Correns rend hommage à Mendel en publiant les deux lois de Mendel :

Première loi de Mendel, ou loi de pureté des gamètes : au moment de la formation des gamètes, il y a ségrégation (*i.e.* séparation) des unités distinctes qui portent les caractères différentiels, qui se réassocient au moment de la fécondation.

Seconde loi de Mendel ou loi de ségrégation indépendante des caractères : les deux unités d'un même couple de caractères ségrégent indépendamment de celles des autres couples.

Dans les années 1910, le généticien Morgan remarque des recombinaisons inattendues lorsqu'il croise des drosophiles portant des caractères différents, portés sur le chromosome X et donc liés physiquement. Les produits issus de ces croisements ont conduit Morgan à remettre en question la seconde loi de Mendel. De plus, un cytologiste Belge, Janssens en 1909, a émis l'hypothèse que les chromatides pourraient se briser et se recoller en réassociant des segments homologues paternels et maternels, conduisant à une recombinaison. Morgan a proposé cette hypothèse pour expliquer ses résultats obtenus et il a nommé « crossing-over » cet échange. La preuve matérielle de ces échanges sera apportée 10 ans plus tard par différents travaux expérimentaux (Haldane, 1934).

A la suite de ces différents travaux, des cartes génétiques, aussi appelées « cartes de liaison » ont été créées. Elles permettent d'observer la distribution des crossing-overs entre les marqueurs qu'elles contiennent. Les cartes définissent l'ordre dans lequel se répartissent les marqueurs et la distance relative qu'il y a entre eux, distance qui correspond à la fréquence de la recombinaison méiotique qui a lieu entre ces marqueurs. On peut donc définir cette distance génétique en unités de recombinaison. Cependant, il n'est pas si évident de construire des cartes génétiques, car les distances exprimées en unités de recombinaison ne sont pas additives. Certes, lorsque la distance physique entre deux marqueurs est suffisamment petite, il ne peut y avoir au plus qu'un seul crossing-over, et dans ce cas la fréquence de la recombinaison est correctement estimée, en revanche, lorsque la distance physique est telle que deux crossing-overs peuvent avoir lieu, les distances génétiques sont sous-estimées. En raison de ce biais, il est préférable d'estimer la distance entre deux marqueurs éloignés comme étant une somme de distances entre des marqueurs intermédiaires. C'est ainsi que, pour pallier cette non-additivité des distances génétiques, Haldane a introduit une distance génétique additive exprimée en centi-Morgan (cM) (Haldane, 1919). Cette distance est une fonction du taux de recombinaison, puisqu'elle est d'autant plus grande lorsque le taux de recombinaison augmente. Elle pourrait s'écrire :

$\mathbf{d = f(R)}$ où $\boldsymbol{d}$ est la distance génétique et $\boldsymbol{R}$ le taux de recombinaison.

La fonction additive $\boldsymbol{f(R)}$ recherchée entre deux marqueurs $\boldsymbol{X}$ et $\boldsymbol{Y}$ est du type :

$\mathbf{d_{XY} = k*\log(1 - 2\ R_{XY})}$ où $\boldsymbol{k}$ est une constante qui doit tenir compte des conditions particulières au voisinage de R = 0.

Lorsque les distances entre les marqueurs $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont très petites, les taux de recombinaison sont très faibles et donc relativement additifs. Ainsi, au voisinage de R = 0, la distance génétique $\boldsymbol{d}$ est égale à $\boldsymbol{R}$, et la fonction peut s'écrire :

$\mathbf{d_{XY} = -2kR}$, d'où :

$\mathbf{k = -1/2}$.

La fonction de distance génétique additive de Haldane s'écrit donc :

$\mathbf{d = (-1/2)*\log(1 - 2R)}$.

Les cartes génétiques ont tout d'abord été utilisées pour ordonner les gènes, puis les marqueurs, sur les chromosomes. En effet, l'estimation de la recombinaison entre les marqueurs

permet de les positionner sur le génome. La création de cartes génétiques précises est essentielle, notamment pour permettre la recherche de *QTLs*. La première carte génétique, construite au tout début du 20ème siècle pour la drosophile, a utilisé des gènes comme marqueurs (Sturtevant, 1913). Cependant, cette approche est limitée, car les scientifiques se sont rendus compte que de nombreux phénotypes étaient affectés par plusieurs gènes polymorphes. Aujourd'hui, les marqueurs les plus utilisés sont appelés des « marqueurs de l'*ADN* ». Ils doivent au moins être bialléliques pour être utiles (Brown, 2002). Quatre types de marqueurs répondent à cette condition : les *RFLP* (Restriction Fragment Length Polymorphisms), les microsatellites, les *SSLP* (Simple Sequence Length Polymorphisms) et les *SNPs* (Single Nucleotide Polymorphisms).

Les *SNPs* sont les marqueurs les plus utilisés aujourd'hui. Ils correspondent à des mutations ponctuelles d'un nucléotide, sont très nombreux tout au long du génome et donc permettent de le baliser relativement finement. La résolution dépend du nombre de marqueurs et du nombre d'individus utilisés pour la réalisation de la carte.

Les marqueurs des cartes génétiques peuvent également être utilisés sur des cartes physiques, ce qui permet de localiser les crossing-overs sur les chromosomes. La carte physique la plus précise correspond à la séquence d'*ADN*, car la distance physique entre chaque *SNP* est alors déterminée à la base près. Ces cartes génétiques et le génotypage d'individus pour ces marqueurs sont aussi utilisés pour étudier la recombinaison. Il est ainsi possible de créer des cartes de recombinaison qui permettent d'étudier la recombinaison dans une population, sachant l'ordre des marqueurs.

L'analyse des pédigrées : méthode familiale

L'analyse des pédigrées permet de créer des cartes génétiques à partir des génotypages des parents et des descendants au sein de familles, de plus ou moins grande taille. Le taux de recombinaison méiotique peut donc être estimé en examinant les transmissions d'allèles à au moins deux loci informatifs spécifiques entre un parent et ses descendants (Capilla *et al.*, 2016). Ceci a permis d'établir plusieurs cartes génétiques individuelles avec des résolutions, c'est-à-dire des densités de marqueurs, différentes et d'observer une variation inter-individuelle du taux de recombinaison. La résolution est cependant limitée par la densité de polymorphismes, par l'informativité des marqueurs, c'est-à-dire par leur nombre d'allèles et par leur fréquence

allélique, et par le nombre de méioses analysées. En revanche, cette méthode permet d'obtenir des taux de recombinaison individuels. Aujourd'hui, les cartes les plus récentes ont une résolution de l'ordre de 10 à 200 Kb, ce qui n'est pas assez précis pour la détection des points chauds de recombinaison (Baudat *et al.*, 2013).

A l'aide de ces pédigrées, il est possible d'identifier des crossing-overs et donc des taux de recombinaison méiotique pour chaque individu étudié. Pour cela, il faut tout d'abord phaser les marqueurs, c'est-à-dire reconstituer la succession des allèles sur les chromosomes transmis à la descendance, selon l'ordre donné par les parents. Les marqueurs homozygotes sont facilement phasés, en revanche, pour les marqueurs hétérozygotes, plusieurs étapes peuvent parfois être nécessaires et différents critères peuvent être utilisés.

Les lois Mendéliennes permettent d'assigner les marqueurs hétérozygotes des descendants à une origine paternelle ou maternelle et seuls ceux pour lesquels les parents sont aussi hétérozygotes ne sont pas phasés (Druet et Georges, 2010). Afin de phaser ces derniers, l'information issue des marqueurs flanquants est utilisée et la probabilité que le marqueur non phasé ait une origine paternelle ou maternelle est calculée. Le marqueur est estimé phasé lorsque l'une des deux probabilités est supérieure à 0,99 (Druet et George, 2010). Il est également possible d'utiliser les informations pédigrées pour le phasage des *SNPs* hétérozygotes chez les parents (Sandor *et al.*, 2012). En effet, des marqueurs « ancres » vont être définis pour lesquels les origines parentales sont connues. Lorsque les parents n'ont pas eux-mêmes leurs parents connus, il n'est pas possible de définir ces marqueurs « ancres », ces derniers seront donc déterminés à partir des descendants pour lesquels l'origine parentale du marqueur peut être déterminée sans ambiguïté (Druet et Georges, 2010). Pour les marqueurs non phasés restants, l'information de ces marqueurs « ancres » va être utilisée et permettra de calculer, là-encore, la probabilité qu'ils proviennent du père ou de la mère. Enfin, une dernière méthode pour aider au phasage des marqueurs, est l'utilisation du déséquilibre de liaison qui se base sur des modèles de Markov cachés (*HMM*) (Druet et Georges, 2010). Le modèle de Markov, aussi appelé Chaîne de Markov, est un modèle statistique composé d'états et de transitions. Une transition matérialise la possibilité de passer d'un état à un autre et dans le modèle de Markov, ces transitions sont unidirectionnelles. Un modèle de Markov caché est basé sur un modèle de Markov, sauf qu'on ne peut pas observer directement la séquence d'états : les états sont cachés. Chaque état émet des

"observations" qui, elles, sont observables. Il n'y a donc pas de travail sur la séquence d'états, mais sur la séquence d'observations générées par les états. Dans le cas des modèles *HMM* utilisés pour modéliser le déséquilibre de liaison, les états cachés (les phases) peuvent être caractérisés par leur fréquence dans la population, par leur fréquence allélique spécifique à la position de chaque marqueur et par des taux de recombinaison calculés pour des intervalles de marqueurs (Druet et Georges, 2010). Le déséquilibre de liaison correspond à une association préférentielle entre deux allèles qui sont donc transmis ensemble à la descendance. Il est souvent utilisé afin d'estimer des taux de recombinaison historique, contrairement aux recombinaisons observées entre un parent et son descendant qui servent à estimer des taux de recombinaison méiotique. Il est possible de n'utiliser que les informations familiales (ségrégation Mendélienne et utilisation des pédigrées) ou que les informations de déséquilibre de liaison, cependant l'association des deux permet d'améliorer la reconstruction des phases (Druet et Georges, 2010).

Afin d'éviter des biais au cours du phasage, les parents choisis doivent, soit avoir leurs propres parents génotypés (Sandor *et al.*, 2012), soit avoir plusieurs descendants, au moins deux chez l'Homme (Coop *et al.*, 2008). A la suite de cette étape de phasage, les crossing-overs sont identifiés comme des changements de phase entre les descendants et les parents (**voir Figure 4**). Le taux de recombinaison individuel correspond ensuite au nombre moyen de crossing-overs par méiose (*i.e.* par descendant).

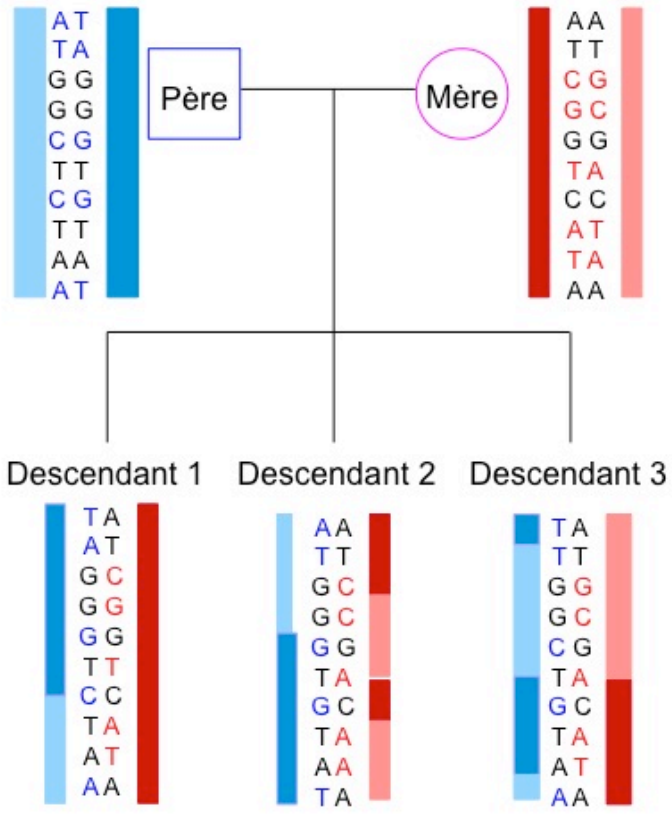

Figure 4: Identification des crossing-overs (d'après Chowdhury et al., 2009)

Observation des génotypes à 10 marqueurs consécutifs pour deux parents et leurs trois enfants. Les marqueurs informatifs, c'est-à-dire ceux pour lesquels un parent est homozygote et l'autre hétérozygote, sont marqués en bleu pour le père et en rouge pour la mère. Les crossing-overs sont identifiés chez les descendants par des changements de couleur (par exemple de bleu foncé à bleu clair).

Chez certaines espèces, il est également possible de créer des cartes fines de recombinaison qui permettent de connaître les taux de recombinaison individuels, mais également la distribution des crossing-overs sur le génome. Cependant, cette dernière observation n'est réellement possible que lorsque les génomes sont de très petites tailles, comme chez le ver ou la drosophile ; cela permet d'avoir un génotypage très dense et donc une très bonne résolution des crossing-overs (Cirulli *et al.,* 2007; Stevison and Noor, 2010; Rockman and Kruglyak, 2009). Pour autant, avec l'évolution à la baisse des coûts de séquençage de génome entier, il sera possible de générer des cartes fines de recombinaison chez les Mammifères par séquençage familial de génome entier, ce qui a déjà été fait chez l'oiseau (Smeds *et al.*, 2016).

L'étude du déséquilibre de liaison : méthode populationnelle

Une autre méthode d'étude génétique de la recombinaison est l'utilisation d'informations sur la diversité génétique de la population, grâce à des méthodes basées sur le déséquilibre de liaison et la diversité haplotypique. Un haplotype correspond à une combinaison d'allèles à plusieurs loci consécutifs (Buard et de Massy, 2007). Les mutations permettent de créer différentes combinaisons d'allèles sur un haplotype et la recombinaison permet de les réassembler et de les mélanger. Le niveau d'association entre les allèles dans la population contemporaine apporte des informations sur les taux de recombinaison historiques.

Le test dit « des quatre gamètes » permet d'estimer si les quatre haplotypes possibles entre deux *SNPs* bialléliques sont présents dans la population. En effet, si dans une population les 4 haplotypes sont observés, c'est qu'il y a eu une recombinaison. Ainsi entre deux *SNPs* consécutifs, si des évènements de recombinaison ont lieu de manière assez fréquente pour pouvoir remanier complètement les allèles, la fréquence ***(ab)*** d'un haplotype donné ne sera pas très différente du produit des fréquences de chaque allèle, ***a*** et ***b*** (Buard et de Massy, 2007). Le déséquilibre de liaison (*DL*), représenté par le coefficient $D\ [D_{ab} = f(ab) - f(a)*f(b)]$, permet de mesurer le degré d'association entre les allèles de deux loci consécutifs. Cependant, le coefficient normalisé ***D'*** est le plus souvent utilisé pour mesurer le déséquilibre de liaison le long du génome sans tenir compte de la fréquence des allèles. Il est égal au rapport D/D_{max} où D_{max} correspond à la valeur maximale (en valeur absolue) que prend ***D*** pour les fréquences observées des gènes. Il peut varier entre 0 et 1.

L'utilisation de ce paramètre a permis de constater que les *SNPs* adjacents ont tendance à former des « blocs », d'environ 10 à 100 Kb de long chez l'Homme (Paigen et Petkov, 2010). A l'intérieur de ces « blocs », tous les marqueurs sont en déséquilibre de liaison. De part et d'autre de ces régions à fort *DL*, il existe des régions plus courtes, de moins de 5 Kb, qui ont un *DL* beaucoup plus faible (Buard et de Massy, 2007). Cette structure « en bloc » existe chez plusieurs Mammifères tels que l'Homme, le chien, le chimpanzé, la souris ou le rat (Buard et de Massy, 2007). Bien que ces motifs de *DL* puissent être influencés par d'autres facteurs, comme les mutations, la recombinaison, la sélection, la démographie ou encore la dérive génétique[3], des

3 La dérive génétique correspond à la modification de la fréquence d'un allèle, ou d'un génotype,

analyses utilisant des méthodes statistiques de coalescence[4] ont été utilisées afin de calculer les probabilités que les frontières entre les « blocs » puissent correspondre à des points chauds de recombinaison, c'est-à-dire de très petites régions du génome (environ 2 Kb) où s'accumule la recombinaison (Paigen et Petkov, 2010).

L'analyse du déséquilibre de liaison s'appuie sur le génotypage d'individus non apparentés. Il est possible d'obtenir la fréquence moyenne de crossing-overs par génération et il est possible de détecter des points chauds de recombinaison. Il n'est cependant pas possible de déterminer des cartes individuelles avec cette méthode ; l'information est uniquement populationnelle et permet de détecter des évènements de recombinaison historiques (Capilla *et al.*, 2016). Plusieurs méthodes existent pour étudier la recombinaison populationnelle, notamment la méthode de LDhat (McVean *et al.*, 2004), qui est la plus utilisée et qui s'appuie sur l'estimation de vraisemblances pour estimer les taux de recombinaison populationnels, ou Ldhelmet qui est dérivée de la précédente et qui a surtout été utilisée chez la drosophile (Chan *et al.*, 2012).

Il est également possible d'estimer un taux de recombinaison populationnel global $\boldsymbol{\rho}$, qui est égal à :

$\boldsymbol{\rho = 4N_e c}$ avec :

- $\boldsymbol{N_e}$ correspondant à la taille efficace de la population[5].

---

au sein d'une population, indépendamment des mutations, de la sélection naturelle ou des migrations, du fait de la taille finie des populations.

4 Les méthodes statistiques de coalescence permettent de reconstruire l'histoire de la population en simulant la généalogie des gènes jusqu'à l'ancêtre commun le plus récent de tous les allèles présents à ce jour dans la population (Paigen et Petkov, 2010).

5 La taille efficace N_e d'une population est définie comme étant la taille d'une population « idéale » où la dérive génétique, évolution des fréquences alléliques qui aboutit à la fixation d'un allèle et donc à la perte de variabilité génétique, aurait la même intensité que dans la population (ou bien le modèle de population) qui nous intéresse.

(http://www.afhalifax.ca/magazine/wp-content/sciences/ConservationBiodiversite/modeleWF/ED-2009_Vitalis_Taille_Efficace_Coalescence.pdf).

- ***c*** correspondant au taux de recombinaison méiotique.

De plus, il est possible d'estimer un taux et une intensité spécifique de recombinaison dans un intervalle ***j*** entre deux marqueurs (Li et Stephens, 2003) :

$\boldsymbol{\rho_j = 4N_e\lambda_j c}$ avec :

- $\boldsymbol{\rho_j}$: taux de recombinaison dans l'intervalle ***j***.

- $\boldsymbol{\lambda_j}$: intensité de recombinaison spécifique à l'intervalle ***j***. C'est un paramètre permettant d'estimer à quel point le taux de recombinaison de l'intervalle ***j*** dévie du taux de recombinaison de base $\boldsymbol{\rho_j}$.

Les méthodes populationnelles, bien que majoritairement utilisées pour la détection des points chauds, peuvent également être utilisées uniquement pour estimer un taux de recombinaison populationnel, ce qui permet d'observer une variation de ce taux de recombinaison, ainsi que d'estimer une corrélation avec la recombinaison méiotique, lorsque ces deux recombinaisons sont étudiées au sein d'une même espèce. C'est par exemple le cas pour le ver *C. elegans*, où une corrélation de 69 % a pu être établie entre la recombinaison populationnelle et la recombinaison méiotique (Rockman et Kruglyak, 2009).

II. 1. b. *Facteurs déterminants la distribution des crossing-overs le long du génome*

II. 1. b. a. Distribution des cassures double-brin

Il a été démontré que les *DSBs* n'étaient pas distribuées de manière homogène sur le génome. En effet, chez *S. cerevisiae*, leur distribution a été étudiée, sur le chromosome III d'individus mutants pour le gène *RAD50.* Cette mutation conduit à l'accumulation de *DSBs* non réparées (Baudat et Nicolas, 1997). Or, la grande majorité de ces *DSBs* a lieu dans des zones intergéniques et promotrices de gènes. Cette accumulation de *DSBs* forme ce qu'on pourrait appeler des « points chauds » de cassures double-brins. Certains de ces points chauds sont également retrouvés dans des régions éloignées d'environ 20-120 Kb des télomères, en revanche, ils sont absents sur des régions d'environ 20 Kb autour du centromère (Martinez-Perez et Colaiacova, 2009). Des études ont également précisé que les *DSBs* se formaient préférentiellement

au milieu des boucles de chromatides, donc en dehors de l'élément axial[6] (Blat *et al.*, 2002). Ces observations semblent donc illustrer que la distribution des *DSBs* est à la fois contrôlée localement, par la structure de la chromatine, mais aussi par la présence de secteurs chromosomiques, comme le centromère ou les télomères par exemple (Martinez-Perez et Colaiacova, 2009).

II. 1. b. b. La distribution des crossing-overs est contrôlée au niveau de la séquence génomique

La mise à disposition de plusieurs génomes et l'augmentation des études génomiques permettent de rechercher une potentielle relation entre la distribution des crossing-overs et des éléments génomiques particuliers, tels que la présence/absence de gènes, de séquences répétées, de taux de *GC* ou encore de motifs d'*ADN.*

Relation entre la distribution des crossing-overs et les gènes ?

Les *DSBs*, initiatrices de crossing-overs, se forment majoritairement dans des régions intergéniques et promotrices de gènes. De plus, il semblerait que la distribution des crossing-overs suive plus ou moins celle des gènes. En effet, chez le blé, Qi *et al.* (2004) ont montré que la densité de gènes pour chaque chromosome augmente lorsqu'on s'éloigne du centromère (corrélation de Pearson de 0,57). La relation entre le taux de recombinaison et la densité de gènes a aussi pu être démontrée chez l'Homme par Kong *et al.* (2002) à l'échelle du génome. Elle est également appuyée par le fait que, chez la levure, la recombinaison est liée à la transcription ; elle est beaucoup plus importante lorsque les sites de fixation des facteurs de transcription sont intacts (Petes, 2001). Cependant, lorsque ce potentiel lien est observé à une échelle plus réduite, chez l'Homme par exemple, il apparaît que le taux de recombinaison est en moyenne plus faible au sein des gènes, mais qu'il augmente lorsqu'on s'en éloigne (jusqu'à une distance d'environ 30 Kb), avant de diminuer de nouveau (Myers *et al.*, 2005). Ce qui est confirmé par Coop *et al.* (2008), qui montrent que le taux de recombinaison augmente lorsqu'on s'éloigne d'une dizaine, voire d'une

6 Les éléments axiaux sont des éléments avec une forte teneur en protéines et présents sur les chromosomes qui, en se rapprochant et à l'aide des nodules de recombinaison, formeront le complexe synaptonémal.

centaine de paires de bases des gènes. En effet, chez l'Homme, le chimpanzé et la souris, la recombinaison a tendance à avoir lieu à quelques dizaines de bases de part et d'autre du site d'initiation de la transcription (*TSS* : Transcription Start Sites) (Przeworski, 2016). Par contre, pour des souris KO pour le gène *PRDM9*, initiateur des *DSBs*, la recombinaison a majoritairement lieu au niveau des *TSS*. Ceci est retrouvé chez les chiens et les oiseaux, chez qui *PRDM9* est inactif ou inexistant, et pour lesquels la recombinaison a également lieu au niveau des îlots *CpG* (Przeworski, 2016). Les éléments de séquence *CpG* correspondent à un segment d'*ADN* de deux nucléotides *C* et *G*. La notation « *CpG* » est une abréviation de « cytosine-phosphate-guanine », destinée à être clairement distinguée de la notation « *CG* » qui peut également désigner une paire de bases sur deux brins d'*ADN* distincts, et non la séquence d'un brin d'*ADN* donné. Dans les génomes, ces dinucléotides *CpG* ont une distribution particulière, car ils définissent des îlots *CpG* dans lesquels leur concentration est élevée. Ces îlots jouent un rôle dans la régulation de l'expression génique. Il semblerait donc que le gène *PRDM9* conduise la recombinaison à s'éloigner des gènes. Chez la levure, les *DSBs* se forment au niveau des promoteurs, mais 68 % de certaines zones très riches en crossing-overs recoupent des séquences codantes (Mancera *et al.*, 2008).

Relation entre la distribution des crossing-overs et les séquences répétées ?

Il existe *a priori* un lien entre certaines séquences répétées et le taux de recombinaison. Ainsi, deux rétrotransposons *THE1A* et *THE1B* sont fortement surreprésentés au sein de zones riches en crossing-overs, chez l'Homme. C'est particulièrement vrai pour une courte séquence de 7 nucléotides *CCTCCCT* qui explique 11 % des points chauds (Myers *et al.*, 2005). Le même résultat a pu être déterminé pour de courtes séquences riches en *CT* ou en *GA*. En revanche, les séquences répétées de *TA* semblent être sous-représentées. Cette séquence consensus de 7 nucléotides est également retrouvée chez la poule avec une corrélation de 0,52 avec le taux de recombinaison et un ratio de 1,99 entre les zones à forte concentration en crossing-overs et les zones fortement dépourvues (Groenen *et al.*, 2009). En revanche, les séquences terminales longues répétées (*LTR*) sont quatre fois plus importantes dans les zones pauvres en crossing-overs, ainsi que les longs éléments nucléaires intercalés (*LINE*). De même que chez l'Homme, les séquences répétées de *TA* sont également faiblement corrélées avec le taux de recombinaison (Groenen *et al.*, 2009). Les mêmes résultats sont retrouvés chez le cochon, les corrélations entre le taux de recombinaison et

les séquences *LINEs* et *LTRs* étant négatives (Tortereau *et al.*, 2012). De plus, certaines séquences spécifiques se retrouvent majoritairement dans des zones riches en crossing-overs, notamment la séquence *CCCCACCCC* qui est trois fois plus fréquente que dans des zones pauvres en crossing-overs. Les motifs *CCTCCCT* et *CCCCACCCC* sont également présents dans des zones riches en crossing-overs chez la souris (Shifman *et al.*, 2006).

Simon Meyers et ses collaborateurs en 2008 ont pu préciser cette courte séquence de 7 nucléotides qui est en fait liée à une séquence plus longue de 13 nucléotides : *CCNCCNTNNCCNC* qui est présente dans près de 40 % des zones riches en crossing-overs (points chauds) chez l'Homme. Plus tard, en 2010, il a été montré que ce motif d'*ADN* avait effectivement un rôle très important dans la recombinaison chez l'Homme (Myers *et al.*, 2010). En effet, c'est parce qu'ils sont reconnus par une protéine à doigts de zinc spécifique, *PRDM9*, que les *DSBs* sont initiées et que des crossing-overs peuvent ensuite apparaître.

Relation entre la distribution des crossing-overs et le taux de *GC*, ainsi que les îlots *CpG*

Les îlots *CpG* correspondraient à des régions particulières où la méthylation des *C* (cytosines) serait inhibée dans la lignée germinale (Gardiner-Garden et Frommer, 1987). La méthylation modifie des cytosines par ajout d'un groupement méthyle, ce qui peut inactiver certains gènes. Les cytosines méthylées sont souvent enrichies dans les îlots *CpG*. Les zones riches en *GC* correspondent à des régions où les éléments *GC* représentent 30 à 60 % de la composition en bases totales de ces zones.

Parmi les premières études entre la distribution des crossing-overs sur le génome et des éléments particuliers de séquences, Kong *et al.* (2002) ont pu démontrer, chez l'Homme, que le nombre de crossing-overs était corrélé aux taux de *GC* et aux îlots de *CpG*. Ils ont analysé l'effet, entre autres, du taux de *GC* et des îlots *CpG*, au sein d'une régression et ils ont ainsi pu observer que lorsque ces effets sont corrigés, le taux de *GC* est négativement corrélé avec le taux de recombinaison et les régions avec le taux de recombinaison le plus important sont celles pauvres en *GC*, mais riches en *CpG*. C'est également le cas chez le rat et la souris ; les régions pauvres en *GC* mais riches en îlots *CpG* montrent un taux de recombinaison plus important (Jensen-Seaman *et al.*, 2004, Shifman *et al.*, 2006). Un résultat similaire est observé chez le chien, à la suite d'une régression multiple, le taux de *GC* est corrélé négativement avec le taux de recombinaison, à

l'inverse des îlots *CpG* (Auton *et al.*, 2012). Ces derniers sont plus fortement liés au taux de recombinaison. Précédemment, nous avons remarqué que le taux de recombinaison est plus élevé à proximité des sites de transcription des gènes, or ces zones sont riches en *CpG*, il se pourrait donc que ce soit plus la présence de ces îlots que les sites de transcription eux-mêmes qui soit à l'origine d'une élévation du taux de recombinaison (Auton *et al.*, 2012). En revanche, bien que le lien entre le taux de *GC* et le taux de recombinaison ait également été montré chez le poulet et le cochon, il apparaît que ce serait une corrélation plutôt positive. En effet, il y a une forte corrélation entre ces deux paramètres, ainsi qu'un enrichissement de *GC* dans les zones riches en crossing-overs (Groenen *et al.*, 2009, Tortereau *et al.*, 2012). De plus, il semble que chez les truies la corrélation entre le taux de recombinaison et le taux de *GC* est plus élevée, notamment lorsque le taux de *GC* est supérieur à 40 % (Tortereau *et al.*, 2012).

Il n'est pas encore clairement démontré si c'est le taux de *GC* qui augmente le taux de recombinaison ou si c'est le taux de recombinaison qui conduit à un enrichissement en *GC*. Certains estiment qu'il existerait un mécanisme sous-jacent commun à ces séquences riches en *GC* qui expliquerait l'augmentation du taux de recombinaison. Ainsi, Petes (2001) a émis l'hypothèse que des régions riches en *GC* conduiraient à de possibles modifications des protéines histones, protéines permettant la compaction de l'*ADN*, ce qui induirait la réplication de l'*ADN*, signal qui serait reconnu par les mécanismes de recombinaison. D'autres études, au contraire, ont proposé qu'un fort taux de *GC* soit dû à la recombinaison. L'une des hypothèses expliquant ce phénomène, présentée par Meunier et Duret (2004), est que la recombinaison favorise l'augmentation des allèles *GC* lorsque les deux types d'allèles, *AT* et *GC*, sont présents dans la population à travers le mécanisme de biais de conversion génique, *BGC*. Il correspond à un mécanisme biaisé de réparation des erreurs de l'*ADN*. Le mécanisme étant biaisé en faveur des bases *GC*, les fréquences alléliques vont donc évoluer de manière non neutre et les hétérozygotes *GC/AT* présenteront une plus forte proportion de gamètes *GC* que *AT*. Ceci conduit à une plus grande probabilité de fixation des allèles *GC* dans la descendance (Galtier *et al.*, 2001). Or, les crossing-overs, et en particulier les non-crossing-overs, ont lieu suite à la réparation des *DSBs*. Si cette réparation est biaisée et suit le mécanisme de *BGC*, il y aura donc un enrichissement en *GC* au niveau des zones riches en recombinaison.

II. 1. c. *Conclusion intermédiaire : la distribution des crossing-overs sur le génome*

La recombinaison méiotique, et en particulier la formation des évènements de recombinaison, tels que les crossing-overs, est essentielle pour la bonne constitution des gamètes. Pour cela, il est très important que les crossing-overs soient finement régulés, tant au niveau de leur nombre qu'au niveau de leur distribution. De nombreuses études ont donc été développées dans plusieurs espèces, de tous les règnes, afin de tenter d'en dégager les grands principes. L'une des grandes règles ainsi établie est la présence d'au moins un crossing-over obligatoire sur chacun des chromosomes, pour assurer leur bonne ségrégation. La distribution des crossing-overs sur le génome est donc régulée. Il existe, entre autres, une régulation fine à l'échelle de la séquence, les crossing-overs étant, semble-t-il, corrélés au taux de *GC*, à des motifs particuliers de l'*ADN*, à des séquences répétées ou encore à la densité de gènes. Il semblerait également que la structure et la compaction des chromosomes influent sur la recombinaison. De plus, l'étude de la distribution fine des crossing-overs sur le génome a permis d'observer des zones où les crossing-overs semblent s'accumuler et d'autres, au contraire, qui en semblent beaucoup plus dépourvues.

II. 2. *Etude des cartes de recombinaison au niveau génomique*

L'étude de la recombinaison peut se faire à deux échelles différentes : à l'échelle globale du génome, ce qui permet d'observer la distribution des crossing-overs sur le génome dans différentes espèces, mais également à une échelle locale beaucoup plus fine, où il est possible d'étudier les zones identifiées préalablement dans lesquelles s'accumule la recombinaison.

II. 2. a. *La distribution des crossing-overs sur le génome*

Lorsque les crossing-overs sont étudiés, que ce soit par des techniques cytologiques ou par l'étude des cartes génétiques, il apparaît qu'ils ne se distribuent pas de manière homogène sur le génome. En effet, certaines régions sont plus riches en crossing-overs que d'autres. Ainsi, chez certains Mammifères, tels que le chien, le taux de recombinaison est plus élevé au niveau des extrémités subtélomériques, ce qui traduit une augmentation de la concentration en crossing-overs dans ces régions (Auton *et al.*, 2013). Chez l'Homme on observe une répartition des crossing-overs similaire à celle de la souris ; il y a un taux de recombinaison élevé au niveau des

télomères, qui diminue ensuite très fortement au niveau du centromère et de ses environs (**voir Figure 5**) : on passe ainsi d'un taux de 3 cM/Mb au niveau des télomères à un taux de 0,1 cM/Mb au niveau du centromère (Kong *et al.*, 2002). Chez la vache et le porc le taux de recombinaison est plus élevé au niveau des extrémités subtélomériques, ce qui traduit une augmentation de la concentration en crossing-overs dans ces régions (Sandor *et al.*, 2012, Tortereau *et al.*, 2012) et ce qui est similaire à la répartition observée chez l'Homme. C'est également le cas chez la souris, cependant, contrairement à l'Homme, les chromosomes de la souris sont acrocentriques, c'est-à-dire que le centromère est placé au début du chromosome (aux alentours de 0 Mb), le taux de recombinaison est faible à cet endroit, il augmente ensuite pour être quasiment maximal à l'autre extrémité du chromosome (**voir Figure 5**) (Schifman *et al.*, 2006). Une même distribution est observée pour la vache ; les chromosomes étant acrocentriques, certains présentent une baisse du taux de recombinaison vers le centre (Ma *et al.*, 2015). Il semblerait donc que, pour ces espèces, la recombinaison au niveau des régions centromériques est inhibée. Il y a plusieurs hypothèses pouvant expliquer ce phénomène, la première serait que la cellule empêche la formation de crossing-overs car cela pourrait conduire à une mauvaise ségrégation des chromosomes (Lamb *et al.*, 2005). La deuxième estime que l'association des centromères en paires non-homologues lors de l'initiation de la recombinaison pourrait empêcher la formation de crossing-overs (Stewart et Dawson, 2008). Il était convenu qu'à large échelle, le taux de recombinaison augmentait avec la densité de gènes, or lorsqu'il est étudié à très petite échelle, il semblerait qu'il soit réduit à proximité des gènes, mais plus élevé lorsqu'on se place à une courte distance de la position de départ des gènes (Coop *et al.*, 2008). C'est vrai chez l'Homme, mais ce n'est pas vrai chez le chien, où le taux de recombinaison reste élevé au niveau des *TSS* (Campbell *et al.*, 2016).

Il y a également une distribution différente du taux de recombinaison en fonction des chromosomes. En effet, très souvent le taux de recombinaison est plus important sur les chromosomes les plus courts. Ainsi, chez l'Homme, le taux de recombinaison des chromosomes 21 (taille de 29,97 Mb) et 22 (taille de 31,19 Mb) est deux fois plus important que celui des chromosomes 1 (taille de 282,61 Mb) et 2 (taille de 252,48 Mb) (Kong *et al.*, 2002). De même, chez la souris où il y a une corrélation négative entre la taille du chromosome et le taux moyen de recombinaison ; le chromosome 19, le plus court, ayant un taux environ 1,5 fois plus fort que celui du plus long chromosome, le chromosome 1 (Shifman *et al.*, 2006). L'interférence peut être une

explication de ce phénomène. En effet, afin d'assurer la bonne ségrégation des chromosomes, il est nécessaire qu'au moins un crossing-over ait lieu par bras chromosomique. L'interférence est donc très élevée sur les grands chromosomes, ce qui réduit l'apparition de crossing-overs, en revanche, sa fréquence est moindre sur les petits chromosomes, afin de permettre l'apparition d'au moins un crossing-over. Cette différence dans la fréquence de l'interférence entre les grands et les petits chromosomes pourrait donc être une explication aux taux élevés de recombinaison sur les petits chromosomes, car cela a déjà été montré chez la levure (Kaback *et al.*, 1999).

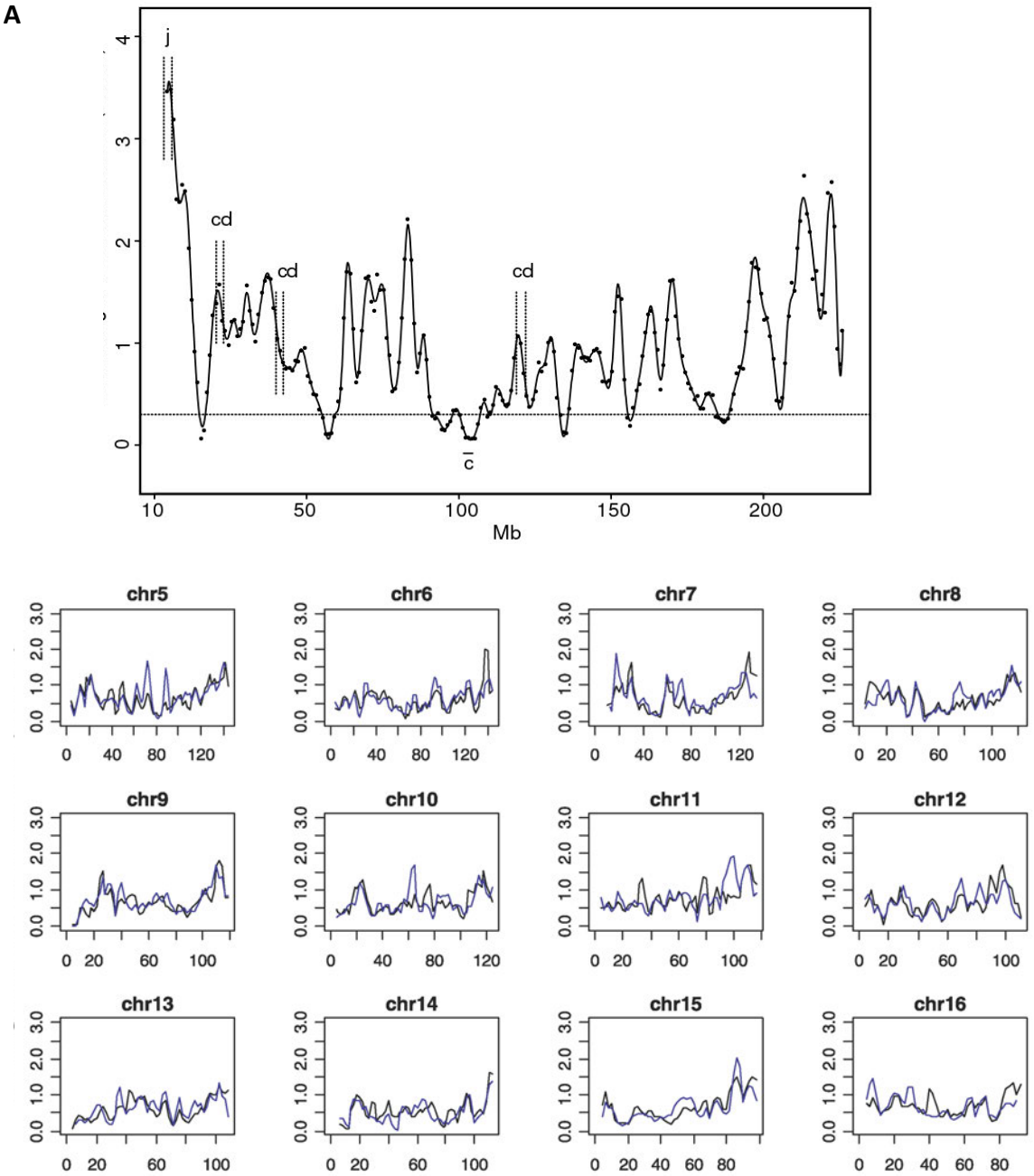

Figure 5 : Répartition du taux de recombinaison dans différentes espèces

A/ Observation de l'évolution du taux de recombinaison moyen chez l'Homme pour le chromosome 3. Les points correspondent au nombre de crossing-overs calculés sur des fenêtres glissantes de 3 Mb (Kong *et al.*, 2002). Le centromère est représenté par la lettre ***c***, les trois lettres ***d*** représentent des régions appelées « désert de recombinaison », la lettre ***j*** représente une « jungle de recombinaison ». Une « jungle de recombinaison » fait référence aux 10 % des intervalles de 1 Mb qui recombinent le plus, et inversement pour les « déserts ». Les chromosomes humains sont métacentriques, le centromère est donc placé au centre. Il est possible de remarquer une élévation du taux de recombinaison au niveau des régions

subtélomériques et une très nette baisse au niveau du centromère. **B/** Observation du taux de recombinaison moyen chez deux lignées de souris (une en noire, l'autre en bleue) pour un échantillon de chromosomes. Le rapport entre la distance génétique et la distance physique a été calculé pour des fenêtres glissantes de 5 Mb (Shifman *et al.*, 2006). La souris ayant des chromosomes acrocentriques, le centromère est placé au début du chromosome et un taux faible est effectivement observé à ce niveau. Taux qui augmente ensuite plus au moins le long du chromosome jusqu'à l'extrémité terminale.

II. 3. *Etude des cartes de recombinaison au niveau local*

L'étude de la recombinaison à l'échelle génomique, que ce soit de manière cytologique ou de manière cartographique, a permis de démontrer que les crossing-overs ne sont pas répartis aléatoirement sur le génome. Au contraire, il semble qu'ils aient plutôt tendance à se regrouper dans des zones spécifiques du génome. Ces régions, de très petites tailles (inférieures à 2 Kb) sont appelées des « points chauds de recombinaison » (hotspots) (Buard et de Massy, 2007). La technique, dite de « sperm-typing », mise au point par Jeffreys et collaborateurs en 1998, permet de caractériser de manière assez précise les points chauds dans les génomes des plantes et des animaux par l'analyse des spermatozoïdes.

II. 3. a. *Qu'est-ce qu'un point chaud de recombinaison ?*

L'activité des points chauds, c'est-à-dire la fréquence d'apparition des crossing-overs, est en moyenne de 0,075 cM, avec un maximum de 0,90 cM. Les sites d'échange ont une taille de 500 à 2 000 paires de bases et forment une ou deux distributions normales, qui peuvent se chevaucher, à partir d'un point central, symbolisant la migration des jonctions de Holliday (Paigen et Petkov, 2010). Ils ont une distance moyenne d'environ 50-100 Kb (Buard et de Massy, 2007). Les points chauds s'opposent aux « points froids » de recombinaison (coldspot) où le nombre de crossing-overs est très faible (inférieur à la moyenne du génome ou à celui des zones à proximité).

Depuis peu, il est également possible de détecter les non-crossing-overs, et il semble qu'ils aient plus particulièrement lieu au centre du point chaud, dans des zones d'une dizaine à une centaine de paires de bases (Paigen et Petkov, 2010). En raison de ces zones de très petite taille, il est souvent difficile de détecter l'intégralité de ces non-crossing-overs, car il y a un manque de marqueurs spécifiques dans ces régions. Ceci rend donc compliqué la comparaison relative de la

proportion de crossing-overs et de non-crossing-overs dans les points chauds.

D'après Petes (Petes, 2001), il existe, *a priori*, trois types de points chauds déterminés par leur activité :

- les points chauds dits de type α : ils seraient dépendants des sites de fixation de facteurs de transcription, mais pas de la transcription elle-même (c'est notamment le cas pour le point chaud de la souris Eβ au locus d'histocompatibilité majeure).

- les points chauds dits de type β : ils seraient associés à des zones dépourvues de nucléosomes, ce qui les rend vulnérables aux nucléases. Ils ont pu être découverts chez la levure, cependant, ils n'ont pas été détectés chez des Mammifères à ce jour.

- les points chauds dits de type γ : ils seraient associés à des séquences riches en *GC*. Ainsi, la séquence poly(*GT*), connue pour stimuler la recombinaison chez la levure, est corrélée avec des régions à forte recombinaison sur le chromosome 22 de l'Homme.

Les points froids ont notamment été étudiés chez *S. cerevisiae*, ils sont surtout présents au niveau des télomères et du centromère et on retrouve une nomenclature semblable à celle des points chauds (Petes, 2001). Les points froids α, correspondent à des domaines chromosomiques où des histones modifiées seraient reconnues par des protéines qui inhiberaient à la fois l'expression des gènes et la recombinaison. A l'opposé de ces régions, où la recombinaison est activement inhibée, les points froids β correspondraient à des zones où il n'y a pas de points chauds. Et donc, contrairement aux points froids α, si un point chaud est inséré au niveau des points froids β, il est possible qu'il y ait une augmentation de la recombinaison. Cependant, il semblerait que, à part au niveau du centromère, il n'y a pas de région du génome de plus de 200 Kb où il n'y a pas du tout de recombinaison (Myers *et al.*, 2005).

II. 3. b. *La détection des points chauds par « sperm-typing »*

La méthode de « sperm-typing » s'appuie sur l'amplification, par *PCR* (Polymerase Chain Reaction), des évènements de recombinaison, notamment des crossing-overs, issus d'un « pool » d'*ADN* extrait des spermatozoïdes. Puis l'analyse des génotypes des recombinants, à l'aide de *SNPs*, permet de reconstruire les phases des différents marqueurs du donneur. Grâce à cette technique, 10 000 000 de spermatozoïdes peuvent être étudiés, et il est possible d'obtenir une

précision de l'estimation du taux de recombinaison d'environ 0,0001 cM à 100 paires de bases et une centaine de recombinants sont connus pour un seul mâle (Buard et de Massy, 2007). De plus, elle permet d'obtenir une carte très détaillée des points chauds avec l'observation à la fois des crossing-overs et des non-crossing-overs (estimation de leur fréquence et de leur structure moléculaire) (Paigen et Petkov, 2010). Cette technique permet d'analyser une quantité importante de cellules provenant d'un même individu, ce qui permet de compenser, notamment chez l'Homme, la difficulté de génotyper de grandes familles (Paigen et Petkov, 2010).

Bien qu'efficace, cette méthode reste cependant très lourde, coûteuse, spécifique des mâles et non adaptée à l'étude de la recombinaison à l'échelle des chromosomes (McVean *et al.*, 2004). Il est en effet difficile d'obtenir un nombre suffisant d'oocytes pour détecter les points chauds, bien que cela ait été réalisé une fois chez une souris femelle (de Boer *et al.*, 2013). Des points chauds ont ainsi pu être étudiés grâce à l'isolation d'*ADN* issu d'une suspension de cellules ovariennes enrichies en oocytes. La méthode de « sperm-typing » surtout utilisée pour étudier des points chauds qui ont déjà été détectés, bien que cette limitation tende à diminuer, étant donné l'importance de plus en plus grande de la caractérisation génomique (séquençage par exemple). La résolution peut également être une limite, car elle dépend de la densité en marqueurs (Capilla *et al.*, 2016).

Il a cependant été possible d'étudier en détail 40 points chauds chez l'Homme et 20 chez la souris (Jeffreys *et al.*, 1998).

II. 3. c. *Les motifs d'ADN caractéristiques des points chauds*

La séquence d'*ADN* présente au niveau d'un point chaud semble primordiale pour la régulation de ce point chaud. Plusieurs études ont cherché à déterminer si les séquences étaient communes aux différents points chauds et donc s'ils étaient contrôlés par les mêmes mécanismes. Ainsi, chez la souris, plusieurs motifs d'*ADN* (éléments répétés et sites de fixation des facteurs de transcription) expliqueraient l'activité des points chauds (Arnheim *et al.*, 2007). C'est chez l'Homme qu'il y a le plus de données disponibles, il est donc possible de rechercher des motifs d'*ADN* qui auraient contribué à l'activité de points chauds. C'est le cas du rétrotransposon *THE1A/B* qui est surreprésenté dans les points chauds (Myers *et al.*, 2005). De plus, lorsqu'il est associé à la séquence *CCTCCCT*, il y a 60 % de chance qu'il y ait effectivement un point chaud de

recombinaison. En revanche, lorsque *THE1A/B* est présent seul, le point chaud n'a plus que 2 à 3 % de chance d'être présent (Myers *et al.*, 2005). Ce motif de 7 bases, comme on l'a vu précédemment, semble donc très important pour le fonctionnement des points chauds. Une étude plus approfondie a permis de préciser ce motif, qui est en fait composé de 13 bases : *CCTCCCTNNCCAC*. La localisation de 41 % des 22 700 points chauds humains est déterminée par la présence de ce motif, de plus lorsqu'il est associé avec un motif *THE1A*, il y a 73 % de chance de détecter un point chaud (Myers *et al.*, 2008). Ces motifs peuvent également être soumis à un polymorphisme qui peut entraîner une différence d'activité au niveau des points chauds. Ainsi, deux points chauds détectés par « sperm-typing », *DNA2* et *NID1*, présentent une transmission biaisée des allèles (par la conversion génique) plus importante chez les individus hétérozygotes pour une position proche du site d'initiation (Buard et de Massy, 2007). De plus, les individus homozygotes pour l'allèle transmis en majorité ont des points chauds moins actifs, ce qui suggère que cet allèle est lié à une diminution de l'activité du point chaud. Or, ce polymorphisme est compris au sein du motif de 7 bases, mentionné précédemment, l'heptamère étant muté pour l'allèle le moins actif : *CC**C**CCCT* au lieu de *CC**T**CCCT* (Buard et de Massy, 2007). Un autre motif de 9 bases *CCCCACCCC*, associé à environ 3 % des points chauds, présente également un polymorphisme associé à une diminution de l'activité des points chauds (Jeffreys et Neumann, 2005). Ces études illustrent à quel point la régulation de l'activité des points chauds est complexe, car, par exemple, au niveau du point chaud *NID1*, les individus hétérozygotes n'ont pas forcément une activité plus importante que celle des homozygotes pour l'allèle diminuant l'activité. Les « jungles de recombinaison », identifiées chez la souris, sont également enrichies avec ces deux motifs (Shifman *et al.*, 2006). Il semblerait qu'il y ait aussi un lien entre les points chauds et les minisatellites hypervariables. Ces derniers pourraient être issus de la réparation des *DSBs* par la recombinaison homologue (Buard et de Massy, 2007). Ainsi, le premier point chaud caractérisé par « sperm-typing » ne se situe qu'à quelques bases d'un de ces minisatellites (Jeffreys *et al.*, 1998). De plus, ils détiennent un fort taux de *GC*, ainsi que des séquences proches des motifs présentés précédemment. Par exemple, chaque unité répétée d'un des minisatellites humain les plus instables, *CEB1*, présente une des trois copies de l'heptamère *CCTCCCT*, concentrant ainsi plus de 100 copies de ce motif sur 2 Kb (Buard et Vergnaud, 1994). Néanmoins, plus de 80 % des points chauds ne sont pas expliqués par des éléments de séquence (Buard et de Massy, 2007). Les

séquences seules, ne doivent pas pouvoir contrôler l'activité des points chauds, car elles sont également présentes dans des régions ne montrant aucune activité de points chauds. Un élément extérieur doit donc sûrement entrer en jeu, or ce motif est reconnu par une protéine à doigts de zinc (Myers *et al.*, 2008). Il a été montré en 2010, chez l'Homme, que c'était la protéine *PRDM9* qui reconnaissait ces motifs d'*ADN*, puis s'y fixait, ce qui enclenchait l'initiation des *DSBs* (Myers *et al.*, 2010). Puis, son rôle chez la souris a ensuite été démontré (Parvanov *et al.*, 2010). En effet, des études préliminaires ont permis d'isoler des loci du chromosome 17, *Dsbc1* et *Rcr1*, qui contrôlaient l'activation de points chauds spécifiques de manière *trans*. L'étude approfondie de la région a montré que dans ce locus se trouvait le gène *PRDM9*, contenant notamment un domaine responsable de la triméthylation de l'histone 3 lysine-4. Etant donné qu'il était exprimé uniquement au début de la méiose, qu'il était capable de se fixer sur des motifs spécifiques d'*ADN* et que l'étude de certains points chauds avait révélé une triméthylation de l'histone 3 lysine-4, le gène *PRDM9* a été considéré comme le candidat idéal pour le locus *Dsbc1/Rcr1*.

La régulation des points chauds ne se fait donc pas seulement à une échelle locale et, de plus, il y a une évolution rapide de leur activité au sein des populations. Ce phénomène s'observe également lorsque les points chauds sont comparés entre les Hommes et les primates, avec lesquels nous pouvons avoir en commun jusqu'à 99 % de notre génome. Bien que des points chauds aient effectivement été détectés chez le chimpanzé, seulement 8 % d'entre eux étaient partagés avec l'Homme (Buard et de Massy, 2007). En outre, la corrélation entre l'activité de la recombinaison des chimpanzés et celle des humains reste très faible, même lorsqu'elle est étudiée sur des fenêtres de plus de 50 Kb. Ceci indique que la distribution des points chauds évolue plus vite que la séquence nucléotidique et c'est l'évolution rapide de *PRDM9*, et notamment de ses doigts de zinc responsables de la reconnaissance des motifs d'*ADN*, qui explique l'évolution des points chauds et leur non conservation entre espèce (Myers *et al.*, 2010).

II. 3. d. *Distribution des points chauds sur le génome*

Les points chauds, tout comme les crossing-overs, ne sont pas distribués de manière homogène sur le génome. Il a été notamment démontré que, chez l'Homme, 80 % des crossing-overs ont lieu dans seulement 10 à 20 % du génome (Buard et de Massy, 2007). C'est également vrai chez le chien, où 80% de la recombinaison a lieu dans 18 à 20% du génome (Campbell *et al.*,

2016). De plus la distribution des points chauds semble coïncider avec la distribution des crossing-overs sur le génome : chez les hommes, par rapport aux femmes, les crossing-overs sont majoritaires dans les régions subtélomériques. Or, il se trouve que les *DSBs* à l'origine des points chauds sont plus denses au niveau des parties les plus distales des chromosomes sur environ 10 Mb, peu importe la taille du chromosome (Pratto et al, 2014). Cette accumulation de crossing-overs à proximité des télomères pourrait être due à la présence de points chauds avec une activité suffisamment forte pour biaiser le rapport *CO/NCO* (Crossing-over et Non-Crossing-Over) en faveur des crossing-overs. Cependant, il est nécessaire d'approfondir ces études afin de pouvoir réellement établir, ou non, une relation entre la proximité des télomères et l'augmentation du ratio *CO/NCO* (Pratto *et al.*, 2014). Le taux de recombinaison élevé observé à proximité des télomères est *a priori* dû une augmentation du nombre de points chauds et non à leur efficacité. Ainsi, il apparaît que dans les régions sub-télomériques, la distance moyenne entre les points chauds est de 90 Kb (contre 123 à proximité du centromère), avec une intensité moyenne d'environ 0,115 cM (0,070 au centromère) (Buard et de Massy, 2007). Cette distribution non aléatoire des points chauds, et notamment leur accumulation aux extrémités des chromosomes, peut être due à l'interférence. En effet, nous avons montré que l'interférence tendait à repousser les crossing-overs les uns des autres. Si la recombinaison avait majoritairement lieu au milieu du chromosome, l'interférence ferait qu'un seul crossing-over pourrait se former. Pour avoir des chromosomes avec 2, voire 3 crossing-overs, une hypothèse possible serait que ces derniers se forment à une extrémité du chromosome, ce qui repoussera le suivant à l'autre extrémité. Ceci pourrait expliquer pourquoi les points chauds sont donc majoritaires aux extrémités des chromosomes.

II. 3. e. *Exemples de points chauds détectés chez les Mammifères*

Les différentes méthodes de détection des points chauds évoquées précédemment ont permis de caractériser plusieurs points chauds dans différentes espèces, notamment chez les Mammifères. Le premier a été découvert en 1982 dans la région *H2* du chromosome 17 de la souris (Steinmetz *et al.*, 1982), résultat d'une recherche approfondie essayant de montrer le lien entre des informations génétiques et la structure moléculaire d'une région chromosomique spécifique. En réalité, des expériences suivantes ont permis de démontrer que cette région (*HS22*)

contenait au moins quatre points chauds. Quelques années plus tard, grâce à l'étude d'analyses pédigrées et du déséquilibre de liaison, le premier point chaud « humain » est détecté dans les régions de la β-globuline et de l'insuline (Paigen et Petkov, 2010). La méthode de détection par « sperm-typing » a permis, comme nous avons pu le voir, de détecter plusieurs points chauds, et notamment la région du complexe majeur d'histocompatibilité (*MHC*), ainsi que celle de la région pseudo-autosomale (*PAR*) des chromosomes sexuels, chez l'Homme (Paigen et Petkov, 2010).

Une base de données appelée « HumHot » a été mise en place afin de répertorier l'ensemble des points chauds connus chez l'Homme (Nishant *et al.*, 2006). Ainsi, environ 25 000 points chauds ont été détectés chez l'Homme, bien que le nombre total soit certainement plus proche de 50 000 ; les 25 000 manquants étant certainement dus aux limites de résolution des puces à *SNPs*, des analyses statistiques ou encore aux limites de la définition des points chauds (Buard et de Massy, 2007). Il a été découvert un nombre similaire chez la souris, entre 15 000 et 20 000 (Baudat *et al.*, 2013). Chez le chien, en revanche, un nombre beaucoup plus limité, 7 677, de points chauds a pu être estimé, avec une largeur moyenne des pics d'environ 4,3 Kb (Auton *et al.*, 2013). Des points chauds ont aussi été détectés chez la vache ; 3 677 sont communs aux deux sexes (Ma *et al.*, 2015). Ces points chauds sont plus densément répartis dans les régions subtélomériques (Ma *et al.*, 2015).

Au vu de ces différentes analyses, il apparaît donc que les points chauds sont un élément très conservé au sein des Mammifères. Il est donc possible de se demander s'ils sont partagés entre les espèces et s'ils sont régis par les mêmes mécanismes. En revanche, pour pouvoir réellement comparer le nombre de points chauds entre les différentes espèces, il faudrait qu'ils aient tous été détectés avec une même méthode, afin d'éviter les biais dus aux méthodes. Les nombres indiqués ici sont donc surtout donnés à titre informatif.

II. 3. f. *Conclusion intermédiaire : les points chauds*

Les points chauds de recombinaison correspondent à de très petites zones du génome, où la concentration en crossing-overs est beaucoup plus importante qu'ailleurs. Ils n'ont pu être découverts que relativement récemment grâce à différentes techniques, aussi bien moléculaires, que quantitatives et mathématiques avec l'utilisation du déséquilibre de liaison. Ils semblent être un élément fondamental de la recombinaison, car présents chez de nombreux organismes.

Les points chauds se distribuent de la même manière que les crossing-overs sur le génome, se localisant majoritairement au niveau des régions télomériques et certaines séquences d'*ADN* semblent plus particulièrement les caractériser. C'est notamment le cas de deux motifs particuliers de 13 et 9 bases qui se retrouvent chez de nombreux points chauds humains, ainsi que dans des « jungles de recombinaison » de la souris et qui sont reconnus par la protéine à doigts de zinc *PRDM9*. Cependant, ces séquences ne peuvent expliquer à elles seules la régulation des points chauds, d'autant plus qu'elles sont soumises à du polymorphisme lié à une évolution rapide des points chauds.

III. Variation inter-individuelle de la recombinaison

Il existe deux phénotypes de la recombinaison qui sont communément étudiés. Il s'agit du phénotype du taux global de recombinaison et du phénotype de la localisation de la recombinaison. De précédentes études ont montré que ces phénotypes pouvaient varier entre les individus et même au sein des individus. L'étude de ces phénotypes se fait majoritairement à l'échelle de la population, il est donc nécessaire d'avoir un grand nombre de données : beaucoup de méioses, de crossing-overs, ou encore une grande densité de marqueurs génotypés.

III. 1. *Le taux global de recombinaison*

III. 1. a. *La mesure du taux global de recombinaison*

Cette mesure peut se faire de deux manières : par étude cytologique, en comptant le nombre de foci de *MLH1* par exemple (Lyrakou *et al.*, 2007), ce qui donne comme phénotype, pour un individu donné, le nombre de crossing-overs par cellule. La mesure peut également se faire de manière indirecte. Dans ce cas-là, les crossing-overs sont identifiés à la suite de la reconstruction des phases entre un parent et ses descendants (cf. « II. 1. a. b. »), le phénotype obtenu pour un individu donné, correspond, cette-fois, au nombre de crossing-overs identifiés par méiose.

III. 1. b. *Mise en évidence d'une variation du taux de recombinaison entre les individus*

Le processus de recombinaison est très conservé entre les espèces, qui présentent un taux de recombinaison global plus ou moins fort. Cependant, plusieurs études ont montré que ce taux pouvait varier entre les individus, notamment chez les Mammifères. Chez l'Homme, ce taux est effectivement variable entre les individus : entre 25 et 27 crossing-overs par méiose (Coop *et al.*, 2008, Chowdhury *et al.*, 2009). C'est également le cas chez la souris, ce qui a été mis en évidence par des analyses cytologiques, avec l'étude de la protéine *MLH1*, qui illustre bien les évènements de recombinaison, et notamment les crossing-overs. Et ainsi, il apparaît que le nombre de foci varie entre 19 et 32 foci par méiose chez la souris, soit une différence de près de 6 crossing-overs entre les individus (Koehler *et al.*, 2002). Cette variation est également présente chez le chien (de 16 à 29 crossing-overs par méiose) (Wong *et al.*, 2010). Il y a également une variation inter-individuelle du taux de recombinaison chez les animaux d'élevage. Ainsi, chez la vache, le taux de recombinaison sur l'ensemble du génome varie entre 18,7 et 32,1 crossing-overs par méiose (Sandor *et al.*, 2012). Une étude similaire a été faite chez le porc et a permis de montrer une variation de 30,2 à 32,4 foci par méiose, soit une différence de l'ordre de 1 crossing-over. De plus, une variation intra-individu a également été démontrée, avec, par exemple en race porcine Large White, un nombre de foci par cellule (spermatocyte) qui varie de 21 à 32 (Mary *et al.*, 2014). Il est intéressant de noter que l'Homme et le cochon présentent la variation de la recombinaison la plus faible entre les individus (de l'ordre de 1 crossing-over par méiose), alors que cette variation est beaucoup plus marquée chez le chien et la vache par exemple (de l'ordre de 13 crossing-overs).

III. 1. c. *Le taux de recombinaison est un phénotype héritable*

Ce phénotype de recombinaison a été étudié dans de nombreuses espèces, notamment chez les Mammifères et il a été montré que la variation inter-individuelle est relativement héritable, c'est-à-dire transmissible à la descendance. Chez la souris, l'héritabilité est proche de 0,46 (Dumont *et al.*, 2009). Cette valeur d'héritabilité est supérieure à celle observée chez l'Homme : 0,13 pour les femmes, 0,08 pour les hommes (Kong *et al.*, 2014). Quant à la vache, l'héritabilité de la variabilité inter-individuelle du taux de recombinaison est estimée à 0,13 chez les mâles (Kadri

et al., 2016).

III. 2. *La localisation de la recombinaison*

III. 2. a. *La mesure de la localisation de la recombinaison*

La mise en évidence d'une variation de la localisation de la recombinaison, et donc de l'usage des points chauds, peut se faire de plusieurs manières. Il est possible d'utiliser une méthode directe avec le « sperm-typing » (Jeffrey *et al.*, 1998). Cependant, pour observer une variation entre les individus, il est nécessaire d'avoir suffisamment de spermatozoïdes provenant de sujets différents. Cette technique très lourde, coûteuse et invasive n'est donc pas la plus adaptée pour réaliser l'étude de ce phénotype tout génome.

En revanche, il est possible d'utiliser des méthodes statistiques indirectes qui se basent sur l'estimation de la probabilité qu'un crossing-over tombe dans un point chaud. Cette étude a notamment été réalisée par l'équipe de Coop *et al.* (2008). Ils ont cherché à estimer la proportion de crossing-overs qui tombaient effectivement dans des points chauds préalablement identifiés chez l'Homme par le Consortium International HapMap (2007). Cette analyse nécessite une très forte résolution et donc une densité en marqueurs très importante (autour de 500 000 *SNPs*).

Ils ont ainsi estimé qu'un crossing-over ***c*** chevauche un point chaud avec la probabilité :

$$\mathbf{P(c\ chevauche\ un\ point\ chaud) = \alpha * 1 + (1 - \alpha)P(c\ chevauche\ un\ point\ chaud\ par\ hasard)}$$

avec :

- α : proportion réelle de *COs* qui a lieu dans les points chauds.

P(c chevauche un point chaud par hasard) : estimée en réalisant 100 distributions aléatoires du taux de recombinaison selon une loi Normale (de moyenne 0 et de déviation standard 200 Kb) et en comptant le nombre de crossing-overs résultant de ces simulations chevauchant un point chaud.

Il est ensuite possible d'estimer la vraisemblance ***V*** de α pour un *CO* ***c*** suivant l'équation suivante :

$$\mathbf{V = \delta_c P(c\ chevauche\ un\ point\ chaud) + (1 - \delta_c) * [1 - P(c\ chevauche\ un\ point\ chaud)]\ (1)}$$

avec :

- δ_c = 1 si c chevauche un point chaud et 0 sinon.

Il est ensuite possible de généraliser ***V*** pour un ensemble de crossing-overs en faisant le produit de l'équation **(1)** sur l'ensemble des *COs* :

$$\mathbf{V = \prod_{c} P(c \text{ chevauche un point chaud}) + (1 - \delta c)*[1 - P(c \text{ chevauche un point chaud})]}$$

Cette estimation prend ainsi en compte l'incertitude sur la localisation des points chauds et sur la localisation des crossing-overs puisque le chevauchement dû au hasard est à la fois influencé par la largeur des points chauds estimés et par la taille des intervalles dans lesquels sont déterminés les crossing-overs (Coop *et al.*, 2008).

Lorsque ce phénotype est connu, il est possible de rechercher une différence d'usage des points chauds par les individus (Coop *et al.*, 2008). Pour cela, un log-ratio de vraisemblances est utilisé, permettant de comparer le log de la vraisemblance maximale d'un modèle, pour lequel chaque individu a une proportion de crossing-overs tombant dans les points chauds différente α_{ind}, au log de la vraisemblance maximale d'un modèle où tous les individus ont la même valeur de α. Et pour s'assurer de la significativité de ce ratio de vraisemblances, des permutations sont réalisées, permettant d'assigner de manière aléatoire le même nombre de crossing-overs à chaque individu et permettant de calculer les ratios.

Lorsque les points chauds ne sont pas aussi précisément connus ; lorsque la résolution n'est pas suffisante pour les détecter, il est possible d'estimer la proportion de crossing-overs qui tombent dans des « fenêtre chaudes » de quelques dizaines de Kb (Sandor *et al.*, 2012). Cette étude est possible car nous avons indiqué précédemment que la recombinaison était effectivement corrélée à la densité en points chauds.

Il est également possible de chercher à caractériser directement les points chauds et leur utilisation en ciblant les motifs reconnus par *PRDM9*. Cela a notamment été fait chez la souris, où l'utilisation de points chauds a pu être comparée *in vivo* et *in vitro* et a révélé que les points chauds n'étaient pas utilisés de la même manière (Walker *et al.*, 2015).

Dans tous les cas, le phénotype de la localisation de la recombinaison est beaucoup moins précis que le phénotype de taux de recombinaison individuel.

III. 2. b. *Observation d'une utilisation différente des points chauds selon les individus*

Il existe une variation inter-individuelle de l'activité des points chauds parmi les individus.

Chez plusieurs hommes, des différences, allant jusqu'à 75 fois plus ou moins d'intensité, ont pu être détectées par « sperm-typing » sur les chromosomes 6, 1 et 21 (Buard et de Massy, 2007). Les mêmes points chauds ne sont également pas utilisés par tous les individus. Par exemple, le point chaud humain *MSMT1* ne présente une activité que pour 11 % d'individus (Arnheim *et al.*, 2007). Une variation entre les populations humaines est aussi observée, lorsque les populations Africaines-Américaines, Européennes-Américaines et Asiatiques-Américaines sont considérées, il est possible de remarquer que 54 % des points chauds sont détectés dans les 3 populations, 18 % des points chauds historiques sont détectés dans une seule population et 36 % ne sont détectés que dans une seule population Africaine-Américaine ou Européenne-Américaine (Arnheim *et al.*, 2007). Cependant, ces résultats sont à relativiser, car il n'est pas évident de savoir s'ils sont réellement population-spécifiques ou dus à des problèmes de puissance statistique. Il semble pourtant exister un certain polymorphisme de la fréquence d'apparition des crossing-overs au niveau des points chauds entre les individus, les variations suivant des modèles variés (Fledel-Alon *et al.*, 2011). Ainsi, chez l'Homme, le point chaud *MSTM1b* est actif chez tous les individus, mais avec des fréquences de crossing-overs différentes (Neumann et Jeffreys, 2006). En revanche, le point chaud *MSTM1a* n'est actif que chez quelques hommes et complètement inerte chez les autres. Ces résultats pourraient suggérer que l'activité des points chauds est contrôlée par des variations locales des séquences d'*ADN*, or certains points chauds sont partagés entre des hommes chez qui la recombinaison est réprimée et d'autres chez qui elle est activée. Cela suggère donc plutôt un contrôle distal et/ou épigénétique, ainsi qu'une évolution rapide des points chauds humains due à la conversion génique biaisée (Neumann et Jeffreys, 2006), ce qui a été confirmé en comparant les séquences des points chauds humains avec ceux des chimpanzés (Myers *et al.*, 2010).

En effet, la résolution des *DSBs* entraîne une conversion génique biaisée en faveur de la chromatide non-initiatrice (Buard et de Massy, 2007). Les modèles et les observations empiriques réalisés ont montré que ce biais de conversion génique conduit majoritairement à la transmission de l'allèle défavorable à la présence d'un point chaud (Boulton *et al.*, 1997). Cette transmission biaisée conduit à une augmentation de la probabilité que l'allèle défavorable se fixe dans la population et mène à un forte réduction, voire à l'élimination, du point chaud. Au cours du temps, il devrait donc y avoir une suppression de tous les points chauds du génome, or ce n'est pas le cas

et Boulton *et al.* (1997) ont nommé ce phénomène le « paradoxe des points chauds ». Il dépend de l'activité du point chaud, ainsi que de la distance entre l'allèle et le centre du point chaud, cette distance affectant la probabilité d'avoir une conversion génique (Buard et de Massy, 2007). Plusieurs mutations à l'origine de ce phénomène ont pu être identifiées chez la levure (Petes, 2001), ainsi que chez l'Homme (Jeffreys et Neumann, 2009). Dans les deux études, un point chaud historique apparu il y a quelques 70 000 ans est décrit. Or, dans ce point chaud, la mutation permettant son activité est très peu transmise, suggérant qu'à terme, il va disparaître (Jeffreys et Neumann, 2009). Cependant, étant donné le rôle majeur des crossing-overs dans la ségrégation des chromosomes, si la formation des crossing-overs diminue dans une région, elle doit être compensée ailleurs sur le même bras chromosomique. Chez la levure, l'activité de sites d'initiation est mise en compétition, suggérant un potentiel mécanisme qui répondrait au paradoxe des points chauds (Fan *et al.*, 1997). Coop et Myers en 2007, ont proposé deux hypothèses à la résolution de ce paradoxe. La première est une sélection directe de zones à fort taux de recombinaison à l'échelle du chromosome. Cette hypothèse semblait très prometteuse étant donné le dogme « un crossing-over obligatoire par bras chromosomique » chez l'Homme, ainsi que chez de nombreux autres organismes. Or, pour les taux de recombinaison humains, une telle sélection ne semble pas avoir d'effet significatif sur un point chaud individuellement (Coop et Myers, 2007). Une autre possibilité serait que les allèles à l'origine de nouveaux points chauds ne seraient, dans un premier temps, pas concernés par la conversion génique. Une telle dynamique s'expliquerait s'il existait une compétition entre des points chauds locaux pour un taux fini de recombinaison. Cette compétition supprimerait initialement l'activité des nouveaux points chauds, ce qui réduirait leur probabilité de conversion génique et permettrait aux fréquences d'allèles créant les points chauds d'augmenter dans la population. Puis, lorsqu'un ancien point chaud dominant un nouveau sera supprimé par la conversion génique, l'allèle du nouveau point chaud aura déjà pu hautement dériver dans la population. Le nouveau point chaud sera ensuite lui-même supprimé de la population, à une vitesse proportionnelle à sa nouvelle activité, par la conversion génique et laissera place à d'autres points chauds, et ainsi de suite (Coop et Myers, 2007).

La résolution de ce mécanisme est donc encore méconnue, cependant, certaines études ont démontré que les rôles essentiels de la recombinaison, tels que la rupture de combinaisons de mutations délétères ou la ségrégation correcte des chromosomes ne suffisent pas à maintenir le

point chaud lorsque c'est l'allèle défavorable qui ségrége (Coop et Myers, 2007). L'hypothèse la plus prometteuse serait l'évolution rapide de *PRDM9*, car la création de nouveaux allèles au niveau de ce gène permettrait la « naissance » de nouveaux points chauds (Baudat *et al.*, 2013). En effet, *PRDM9* est un gène présentant de nombreux polymorphismes chez plusieurs espèces, notamment chez la souris (Baker *et al.*, 2015). Or l'existence d'allèles différents peut conduire à des modifications sur l'activité de *PRDM9* et sur sa reconnaissance des motifs d'*ADN* spécifiques des points chauds. La reconnaissance et la fixation d'un motif par *PRDM9* conduisent, à terme, à son érosion. C'est donc la mutation de *PRDM9* et la création de nouveaux allèles qui permettent au gène de reconnaître d'autres motifs et ainsi d'activer de nouveaux points chauds (Baker *et al.*, 2015).

Tout comme pour la variation du taux de recombinaison, la variation inter-individuelle de la localisation de la recombinaison apparaît également héritable, avec une héritabilité de 23 % chez l'Homme (Fledel-Alon *et al.*, 2011) et de 21% chez la vache (Sandor *et al.*, 2012).

IV. Le déterminisme génétique de la variation inter-individuelle du taux de recombinaison et de la variation inter-individuelle de la localisation de la recombinaison

Il a été montré que les deux phénotypes de recombinaison étudiés : la variation inter-individuelle du taux de recombinaison et la variation inter-individuelle de la localisation de la recombinaison, sont tous les deux héritables. Il semble donc qu'ils soient soumis à un déterminisme génétique ; un ou plusieurs gènes en étant responsables.

IV. 1. *Méthodes de détection des QTLs*

En génétique quantitative, la majorité des études sont réalisées avec des cartes génétiques, des puces à *SNPs*. Une potentielle liaison statistique entre les allèles aux marqueurs, les *SNPs*, et le phénotype des individus est ainsi recherchée. Les variabilités inter-individuelles observées

précédemment permettent d'envisager un contrôle par un ou plusieurs gènes de la recombinaison, c'est pourquoi on recherche majoritairement des *QTLs* (Quantitative Trait Loci), régions chromosomiques, loci, comportant un ou plusieurs gènes pouvant expliquer la variabilité entre les individus. En revanche, étant donné que la détection de *QTLs* retourne une zone chromosomique plus ou moins large, l'information peut être assez imprécise et donc des études complémentaires devront être menées afin de confirmer les gènes d'intérêt et d'identifier de potentielles mutations causales.

IV. 1. a. *Définition et principes de base de la détection de QTLs*

Les premières études *QTLs* ont été réalisées au début du XXème siècle sur le pois par Sax (1923). Il a rapporté une relation significative entre la couleur et la taille des graines de haricot et il a postulé que cette association était due à la présence, à proximité du gène contrôlant la couleur, d'un locus ayant un effet quantitatif sur la taille de la graine. Deux critères sont nécessaires à la détection de *QTLs* : la variation au niveau du phénotype considéré et la présence de marqueurs polymorphiques à proximité du *QTL*. Les marqueurs ont des positions connues et couvrent le génome de manière homogène. La détection d'un *QTL* passe par deux étapes : premièrement, le regroupement de la population phénotypée dans diverses classes en fonction de l'allèle porté par les individus et deuxièmement, l'utilisation d'un test statistique permettant de déterminer si une différence significative peut être observée entre les moyennes phénotypiques de chaque groupe (Georges, 2007). Si ce test dépasse un certain seuil, la présence d'un *QTL* pourra être proposée. En revanche, il est souvent difficile de déterminer la mutation causale expliquant réellement le phénotype car il faut pouvoir effectuer des tests fonctionnels sur les potentielles mutations candidates obtenues et ces tests peuvent être lourds et coûteux.

Il existe plusieurs méthodes pour rechercher une association entre un *QTL* et un phénotype étudié, notamment les analyses de liaison, exploitant le déséquilibre de liaison au sein de grandes familles ou encore les études d'associations pangénomiques (*GWAS*), exploitant, cette fois, le déséquilibre de liaison au niveau de la population.

IV. 1. b. *La détection marqueur par marqueur*

L'utilisation de cette méthode suppose que le *QTL* est confondu avec le marqueur génotypé. Il est alors possible de rechercher une différence significative entre les moyennes phénotypiques de plusieurs groupes d'individus, génotypés au marqueur étudié (Soller *et al.*, 1976). Cette différence permet de quantifier l'effet du *QTL*, qui peut être traduit par un modèle linéaire d'analyse de variance :

$\mathbf{y = \mu + x\beta + e}$ avec :

- **y** : un vecteur de taille n contenant les phénotypes de n individus.
- **μ** : la moyenne des phénotypes.
- **x** : un vecteur contenant un marqueur génotypé.
- **β** : l'effet du marqueur.
- **e** : les erreurs aléatoires de distribution normale.

Lorsque le *QTL* est effectivement confondu avec un marqueur, cette méthode de détection est très puissante, en revanche, lorsque le *QTL* se situe entre deux marqueurs, elle présente deux limitations majeures. Premièrement, l'effet du *QTL* diminue lorsque la distance avec le marqueur augmente. Il sera donc nécessaire d'avoir un plus grand nombre de descendants pour détecter un effet significatif au niveau de ce *QTL*. Deuxièmement, la détection marqueur par marqueur ne donne pas la position exacte du *QTL* par rapport au marqueur. Afin de pallier ce manque de précision, on peut utiliser des analyses de liaison, et notamment la cartographie d'intervalle.

IV. 1. c. *L'utilisation des analyses de liaison*

Les toutes premières études sur la recombinaison génétique ont permis de découvrir que certains caractères phénotypiques étaient transmis à la descendance de façon corrélée. Ce qui permettait de supposer que les gènes responsables de ces caractères étaient localisés sur les mêmes chromosomes et que leur degré de corrélation reflétait leur proximité physique. Les analyses de liaison cherchent donc à analyser la corrélation entre un caractère et des marqueurs génétiques transmis au sein d'une famille ; elles cherchent à montrer si des patrons de ségrégation sont associés à la variation d'un phénotype.

Analyse de liaison au sein de lignées consanguines

Dans les lignées consanguines, les individus F_1 reproducteurs sont obtenus en croisant des individus issus de deux lignées parentales F_{01} et F_{02}, puis les individus F_1 sont soit croisés avec une des deux lignées parentales (phénome de rétro-croisement ou « backcross »), soit ils s'accouplent entre eux pour donner la lignée F_2. A travers ces croisements, du déséquilibre de liaison est créé et, même si les évènements de recombinaison peuvent le diminuer, il est possible de l'exploiter à l'aide de cartes génétiques. De plus, étant donné que les pédigrées sont connus, l'origine parentale de chaque marqueur génotypé est connue. Il est cependant nécessaire d'avoir des familles de grande taille, car la puissance des analyses de liaison est majoritairement affectée par la taille de la population et l'effet du *QTL*, la densité de marqueurs ayant un effet moindre (Darvasi *et al.*, 1993).

Finalement, au sein d'une lignée consanguine, les individus peuvent être regroupés en deux (population « backcross ») ou en trois groupes (population F_2), selon leur génotype à chaque marqueur étudié. Dans le cas d'une population « backcross », la moitié des individus est homozygote, l'autre moitié est hétérozygote, et donc pour chaque marqueur, une différence significative entre les deux groupes indiquera la présence d'un *QTL*.

Cartographie d'intervalle dans des lignées « outbred »

Contrairement aux plantes ou aux espèces modèles, comme la souris, les populations d'élevage sont rarement complètement consanguines, ce qui a des impacts sur la détection de *QTLs* (Lynch et Walsh, 1998). En effet, les marqueurs et les *QTLs* ne sont pas toujours en complet déséquilibre de liaison, ce qui fait que les haplotypes transmis ne sont pas forcément les mêmes entre les parents et les descendants.

Dans les lignées « outbred », les individus sont majoritairement hétérozygotes, il est donc beaucoup plus compliqué de reconstituer les origines parentales. Les méthodes de phasage, les mêmes que celles utilisées pour déterminer le nombre de crossing-overs dans une méiose, permettent de reconstruire ces origines parentales (**voir Figure 4**). Les individus n'étant pas forcément tous apparentés de la même manière, il est possible d'avoir plus de 2 allèles au *QTL* recherché. Il est alors nécessaire de les estimer comme des effets aléatoires : c'est le principe d'analyse en composantes de la variance.

L'analyse de liaison a été largement utilisée ces vingt dernières années et a été très efficace pour détecter des polymorphismes de gènes impactant des caractères mendéliens. En revanche, le faible nombre d'évènements de recombinaison rend peu précise la détection d'un variant causal[7] par un marqueur, cette technique rencontre donc beaucoup moins de succès dans l'identification de mutations causales (Visscher *et al.*, 2012). Il y a plusieurs possibilités pour expliquer ce problème de détection, notamment la pénétrance des variants causaux, qui serait trop faible pour permettre l'identification par co-ségrégation, mais également les interactions entre les gènes, telles que l'épistasie. De plus, ces études de liaison demandent l'analyse d'un grand nombre d'individus afin d'obtenir des résultats statistiques suffisamment significatifs.

IV. 1. d. *L'utilisation des GWAS pour la détection de QTLs*

Risch et Merikangas (1996) sont très certainement les précurseurs de la *GWAS*. Ils ont en effet proposé d'utiliser quelques individus, pas forcément apparentés, au lieu de familles, afin de rechercher des *QTLs* et des gènes candidats : c'est ce qu'ils ont appelé des tests d'association. Ils ont montré que, grâce aux lois de Mendel, si un marqueur est associé au gène d'intérêt, alors il sera transmis à plus de 50% à la descendance.

Les *GWAS* s'appuient sur des co-ségrégations historiques entre les marqueurs et les *QTLs*, qui sont mises en évidence par l'analyse du déséquilibre de liaison.

IV. 1. d. a. *GWAS* : définition, avantages et inconvénients

Pour rechercher une potentielle association, dans une population, entre un gène et un phénotype (le déterminisme génétique), il faudrait connaître l'association de chaque polymorphisme du génome avec le phénotype étudié, à moins d'un séquençage très performant, cela est rarement connu. Cependant, du fait du déséquilibre de liaison élevé entre polymorphismes proches, il est possible d'étudier seulement un échantillon de polymorphismes présents sur le génome. En utilisant les informations génotypiques et phénotypiques, une *GWAS*

---

7 Un variant causal est défini comme une mutation qui contribue à influencer de manière positive ou négative un caractère étudié au sein d'une population (Visscher *et al.*, 2012).

peut ensuite être réalisée et permettra de rechercher une association potentielle entre le phénotype et les marqueurs. Il faudra ensuite identifier un gène candidat, influençant le phénotype, qui serait en déséquilibre de liaison avec les marqueurs. L'analyse d'association est une application directe du modèle linéaire d'analyse de variance présenté précédemment.

Une *GWAS* est donc basée sur le déséquilibre de liaison entre des marqueurs génotypés et des variants causaux non génotypés (Visscher *et al.*, 2012). La puissance de l'analyse, c'est-à-dire la probabilité de rejeter correctement l'hypothèse nulle H_0 (pas d'effet du marqueur sur le phénotype d'intérêt) lorsqu'un *QTL* est réellement présent, dépend de l'effet du *QTL*, de sa fréquence et du déséquilibre de liaison entre le *QTL* et le marqueur associé. Concernant ce dernier critère, il a été montré que, pour avoir la même puissance que celle obtenue si le génotype au *QTL* était connu, la taille de l'échantillon devait être augmentée d'un facteur $1/r^2$, r^2 représentant la valeur du déséquilibre de liaison entre le marqueur et le *QTL* (Pritchard et Przeworski, 2001).

La structure de la population[8] est également très importante, en particulier pour les animaux d'élevage. Pour la prendre en compte, il est donc important d'avoir un nombre suffisant de marqueurs. Ainsi, l'équipe de MacLeod (2008) a estimé que pour une population de 365 vaches génotypées pour 10 000 *SNPs*, il était possible de détecter un *QTL* expliquant 5 % de la variance phénotypique avec une puissance de 37 %. La proportion de variance expliquée par le *QTL* peut s'écrire selon la formule suivante (MacLeod *et al.*, 2008) :

$$\sigma_p = 2p(1 - p)\alpha^2 \text{ avec :}$$

- σ_p : proportion de variance expliquée par le *QTL*.
- **p** : fréquence allélique du *QTL*.
- α^2 : effet du *QTL*.

Ceci montre que la fréquence allélique du *QTL* est également très importante pour permettre la détection du *QTL*, et plus le *QTL* est rare, plus il faudra que son effet soit fort pour

8 Une population est dite structurée lorsqu'elle présente de grandes différences dans ces ancêtres, par exemple des niveaux d'immigration différents ou des groupes d'individus qui ont plus d'ancêtres communs que ce qui est attendu dans une population en panmixie (Astle et Balding, 2009).

pouvoir expliquer une certaine proportion de la variation phénotypique observée.

Etant donné qu'une *GWAS* est l'application directe d'un modèle linéaire, le modèle le plus simple est le suivant (Zhang *et al.*, 2010) :

$$\mathbf{y = \mu + x\beta + a + e} \text{ avec :}$$

- **y :** le phénotype étudié.
- **x :** vecteur contenant le marqueur testé.
- **β :** effet du marqueur.
- **a :** valeur génétique de l'individu qui prend en compte l'apparentement, ***a*** suit une loi Normale de moyenne et d'écart-type : $\boldsymbol{N(0, P\sigma^2 a)}$ où ***P*** est une matrice de parenté génomique ou pédigrée.

Comme nous l'avons vu, les *GWAS* reposent sur le principe d'un déséquilibre de liaison entre les marqueurs et un *QTL*, or l'intensité de ce déséquilibre dépend fortement de la fréquence allélique des marqueurs. Les variants rares (avec une fréquence < 0,01) risquent d'être en faible déséquilibre avec un autre marqueur proche (Visscher *et al.*, 2012). Pour cette raison, les marqueurs sélectionnés pour être sur les puces sont considérés comme communs, avec majoritairement une fréquence allélique supérieure à 0,05. Ainsi, les *GWAS* sont conçues pour repérer des associations avec des marqueurs relativement communs dans la population. De plus, en cas de faible déséquilibre entre le marqueur et le *QTL*, ce dernier pourrait ne pas être détecté. Pour pallier ce problème de faible déséquilibre, il est possible d'utiliser des haplotypes, en particulier dans le cas de mutations récentes (Balding, 2006). Cela peut permettre d'avoir une meilleure connaissance des régions régulatrices, proches du gène candidat, qui auraient une action *cis* sur le génome.

Bien que des centaines de polymorphismes ont été identifiés comme ayant un lien avec les caractères recherchés (chez l'Homme, majoritairement des maladies génétiques), cette technique présente malgré tout quelques limites. Tout d'abord, beaucoup d'études ne prennent pas en compte la structure, la stratification de la population, ce qui implique que la majorité des polymorphismes identifiés n'ont en réalité aucun lien biologique avec les phénotypes étudiés (McClellan et King, 2010). Ce biais de structure peut même conduire à des faux positifs qui sont dus à la création de déséquilibre de liaison entre loci éloignés. Il peut également y avoir des

erreurs d'interprétation, l'association pouvant se manifester avec un gène voisin de celui qui est réellement responsable du phénotype étudié. De plus, les *GWAS* sont limitées par la résolution en marqueurs, la densité de génotypages et le nombre d'individus sur lesquels l'étude est réalisée ; plus il y a d'individus étudiés, plus la détection est améliorée.

IV. 1. d. b. Précision de la détection grâce aux méthodes multi-*QTLs*

Contrairement à la méthode d'association qui étudie l'effet de chaque *SNP*, un par un, sur le phénotype, les méthodes multi-*QTLs* permettent de prendre en compte les effets de plusieurs variants génétiques en même temps, notamment le modèle *BSLMM* (Bayesian Sparse Linear Mixed Model) développé par Zhou *et al.*, (2013). Ce genre de modèles a été utilisé pour prédire les valeurs génétiques en sélection génomique chez les animaux et les plantes, pour prédire les phénotypes de caractères complexes dans les organismes modèles et la vache laitière et, plus récemment, pour cartographier des caractères complexes en modélisant conjointement tous les *SNPs* dans des populations structurées (Zhou *et al.*, 2013).

Le modèle *BSLMM* permet de considérer plusieurs *QTLs*, qui ont tous des effets polygéniques sur les différents marqueurs. Le principe de la méthode est d'avoir pour chaque *SNP* ***l*** un indicateur variable γ_l qui prend la valeur de 1 si le *SNP* est un *QTL* ou 0 sinon. La possibilité qu'un *SNP* soit un *QTL* est mesurée par une probabilité *a posteriori* $\boldsymbol{P(\gamma_l = 1)}$, appelée « probabilité postérieure d'inclusion » (*PIP*). Lorsque ce modèle est utilisé, tous les *SNPs* sont évalués. Les estimations des paramètres du modèle sont réalisées grâce à un algorithme itératif *MCMC*, qui est un algorithme d'estimation d'un modèle : le nombre d'itérations est fixé à 10 millions et l'estimation est réalisée en extrayant des échantillons toutes les 100 itérations. Pour savoir si une région est un *QTL*, une somme glissante des *PIPs* est calculée sur 50 *SNPs* consécutifs. Puisque les distances physiques entre les marqueurs d'une puce 600K sont d'environ 5 Kb, le modèle estime la probabilité qu'il y ait un *QTL* pour des fenêtres chevauchantes de 250 Kb, qui comportent environ 20 *SNPs* (Zhou *et al.*, 2013).

IV. 1. d. c. Amélioration de la détection grâce à l'imputation

Etant donné que la densité de génotypages est un facteur limitant pour l'efficacité des *GWAS*, il est intéressant de pouvoir utiliser des puces haut-débit, avec un grand nombre de

marqueurs génotypés (au moins 500 000 marqueurs). Cependant, de telles puces ont un coût et il n'est donc pas toujours possible de les utiliser pour génotyper un nombre suffisant d'individus. En revanche, elles sont très utiles pour prédire une densité accrue de génotypages d'individus génotypés sur des puces plus petites. C'est ce qu'on appelle l'imputation.

L'imputation repose sur le principe de reconstruction des haplotypes. En effet, il a été montré que, sur de très petites régions (quelques kilobases chez l'Homme), les haplotypes avaient tendance à se regrouper dans des groupes d'haplotypes similaires, du fait de la recombinaison. Ces haplotypes, étant très similaires, varient de la même manière le long du chromosome (Scheet et Stephens, 2006). Ces groupes d'haplotypes, présents dans un panel de référence, vont être utilisés pour prédire les génotypes absents dans un échantillon d'individus, la cohorte. La reconstruction des haplotypes peut être réalisée grâce à des méthodes de maximum de vraisemblance (Scheet et Stephens, 2006). Ici, les auteurs ont développé la méthode FastPHASE qui a ensuite été implémentée dans le logiciel BIMBAM (Servin et Stephens, 2007). FastPHASE s'appuie sur un modèle de Markov caché où ***K*** états cachés sont représentés par les haplotypes communs.

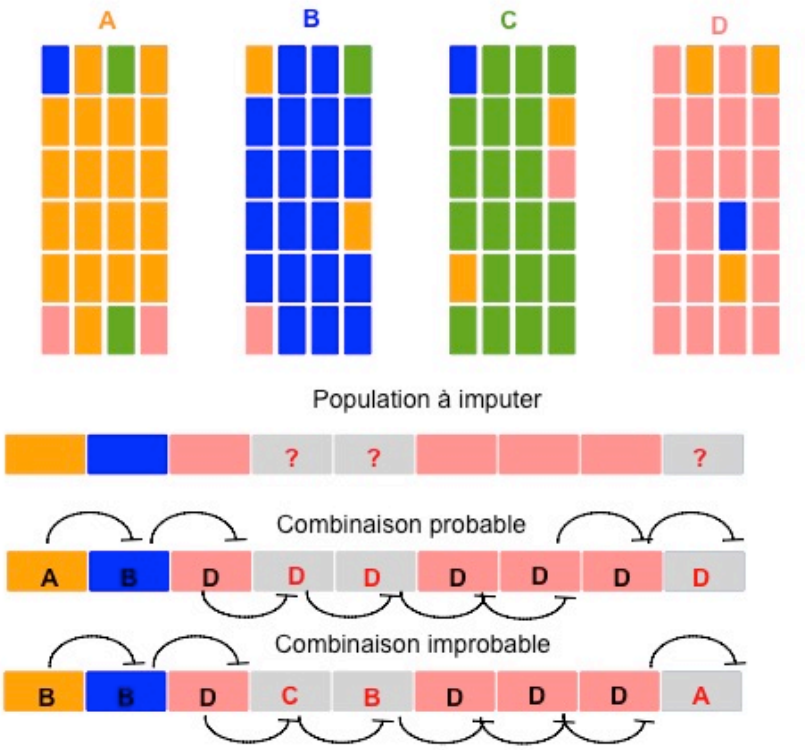

Figure 6: Schématisation de l'imputation

A, B, C et D représentent quatre groupes d'haplotypes utilisés ensuite dans la chaîne de Markov cachée. Les couleurs représentent différents allèles. Dans la population à imputer, les « ? » représentent les marqueurs à imputer et on cherche à savoir quelle combinaison est la plus probable ; de quel groupe d'haplotypes les allèles sont le plus probablement issus.

En résumé, l'imputation utilise deux échantillons : un premier, le panel, avec un faible nombre d'individus, la plupart du temps non apparentés, choisis pour être les plus représentatifs possibles de la population et génotypés sur une puce haute densité. Le deuxième, la cohorte, est constitué des individus sur lesquels la *GWAS* est pratiquée, donc suffisamment nombreux, souvent apparentés et génotypés sur une puce de plus basse densité.

Les modèles du déséquilibre de liaison sont ensuite utilisés afin de prédire les génotypes « manquants » de la puce basse densité, à partir de ceux présents sur la puce haute densité. C'est cette prédiction qui est nommée imputation (**voir Figure 6**). A la fin, un échantillon d'individus à analyser est obtenu, la cohorte, présentant le même nombre de marqueurs que celui de la puce haute densité, cependant, les marqueurs qui ne sont pas « réellement » génotypés chez eux (les

« manquants »), ont été déduits et ont donc un génotype estimé. Il est donc bon, suite à une imputation, d'utiliser des modèles probabilistes qui permettent d'estimer la qualité de cette imputation.

IV. 2. *Le déterminisme génétique de la variation inter-individuelle du taux de recombinaison*

IV. 2. a. *Plusieurs loci associés à ce phénotype*

Plusieurs loci pouvant expliquer la variation du taux de recombinaison ont été découverts (**voir Tableau 2**).

Tableau 2 : Gènes candidats proposés pour la variation inter-individuelle de la recombinaison

Espèce	Locus ou gène candidat	Chr.	Effet	Mécanisme d'action	Référence
Homme	*Inversion 17q21.31*	17	Augmentation de 0,472 Morgans par copie de l'inversion.	Non connu	Stefansson *et al.*, 2005
Homme	*KIAA1462*	10	Augmentation de 1,686 cM/Mb.	Expression dans l'ovaire des souris nouvelles-nées et dans les oocytes en Prophase I.	Chowdhury *et al.*, 2009
Homme	*PDZK1*	1	Augmentation de 2,515 cM/Mb.	Expression au même endroit et au même moment que *KIAA1462*.	Chowdhury *et al.*, 2009
Homme	*UGCG*	9	Augmentation de 1,849 cM/Mb.	Expression maximale au stade Diplotène de la Prophase I.	Chowdhury *et al.*, 2009
Homme	*PCGF3*	4	Augmentation de 124 cM.	Formation d'un complexe avec *GAK* et implication dans les points de contrôle du cycle cellulaire et dans les mécanismes de réparation de	Kong *et al.*, 2014

				l'*ADN*.	
Homme	*SMEK1*	14	Augmentation de 1 cM.	Rôle dans la réparation de l'*ADN*.	Kong *et al.*, 2014
Homme	*RAD21L*	20	Augmentation de 48 cM.	Protéine spécifique de la méiose exprimée dans les spermatocytes et les oocytes.	Kong *et al.*, 2014
Homme	*CCDC43*	17	Diminution de 5 cM.	Non connu.	Kong *et al.*, 2014
Homme/Vache	***RNF212***	**4/6**	**Augmentation de 124 cM. / Augmentation de 3,3 *COs*.**	**Nécessaire à la formation des crossing-overs.**	**Kong *et al.*, 2008, 2014 Chowdhury *et al.*, 2008 Sandor *et al.*, 2012 Ma *et al.*, 2015 Kadri *et al.*, 2016**
Homme	***GAK***	**4**	**Augmentation de 124 cM.**	**Association avec le complexe *Cycline-G* qui participe à la recombinaison chez la Drosophile.**	**Kong *et al.*, 2014 Nagel *et al.*, 2012**
Homme/Vache	***CPLX1***	**4/6**	**Augmentation de 124 cM. / Augmentation de 1,09 cM/Mb.**	**Pas de rôle connu avec la recombinaison.**	**Kong *et al.*, 2014 Ma *et al.*, 2015**
Homme	***CCNB1IP1/HEI 10***	**14**	**Augmentation de 16 cM.**	**Essentiel pour la formation de crossing-overs chez la souris.**	**Kong *et al.*, 2014 Ward *et al.*, 2007**
Vache	***REC8***	**10**	**Différence de 1,8 crossing-overs entre les homozygotes.**	**Code pour une protéine localisée sur le complexe synaptonémal. Elle assure la cohésion entre les chromatides sœur et la**	**Sandor *et al.*, 2012**

				recombinaison entre les homologues.	
Vache	***RNF212B***	**10**	**Augmentation de 1,24 cM/Mb.**	**Paralogue de *RNF212*, mais pas de fonction connue en lien avec la recombinaison.**	**Kadri *et al.*, 2016**
Homme/Vache	*MSH4*	1/23	Diminution de 2 cM. / Augmentation de 0,10 cM/Mb.	Essentiel à la formation des crossing-overs à travers le complexe MutSγ.	Kong *et al.*, 2014 Ma *et al.*, 2015 Kadri *et al.*, 2016
Vache	*MSH5*	23	Diminution de 1,07 cm/Mb.	Rôle lors de la réparation des *DSBs* et formation des crossing-overs.	Kadri *et al.*, 2016
Vache	*SMC3*	26	Augmentation de 0,17 cM/Mb.	Code pour une protéine du complexe synaptonémal.	Ma *et al.*, 2015
Vache	*MLH3*	10	Augmentation de 0,74 cM/Mb.	Impliqué dans la réparation des *DSBs*.	Kadri *et al.*, 2016
Vache	*HFM1*	3	Diminution de -0,65 cm/Mb.	Formation des intermédiaires de recombinaison.	Kadri *et al.*, 2016

Les informations en rouge concernent les données chez l'Homme, tandis que les informations en bleu concernent les données chez la vache. Les bordures vertes et bleues regroupent des gènes qui se trouvent dans des mêmes loci. Les gènes en gras sont ceux les plus retrouvés dans les différentes études.

Malgré l'existence de tous ces gènes, la majorité n'est pas retrouvée quand les différentes études sont comparées, rendant discutable leur influence sur le phénotype. En revanche, un gène a été démontré comme ayant un réel rôle sur la variation du taux de recombinaison, il s'agit de *RNF212*.

IV. 2. b. *Intérêt majeur pour RNF212*

RING finger protein *RNF212*, est un gène candidat quasiment universel chez les Mammifères pour lesquels le déterminisme génétique de la variation inter-individuelle du taux de recombinaison a été étudié. Il a été détecté chez l'Homme, sur le chromosome 4 (Kong *et al.*, 2014, Chowdhury *et al.*, 2009) et chez la vache, sur le chromosome 6 (Sandor *et al.*, 2012, Ma *et al.*, 2015, Kadri *et al.*, 2016). Des polymorphismes du gène ont pu être directement reliés au phénotype chez l'Homme (Chowdhury *et al.*, 2009, Kong *et al.*, 2014) et chez la vache (Kadri *et al.*, 2016), faisant de lui un très bon candidat positionnel et fonctionnel. De plus, les souris mâles mutants pour ce gène sont stériles ; ils ne produisent pas de sperme et ont des testicules beaucoup plus petits que les sauvages (Reynolds *et al.*, 2013).

Son rôle dans la recombinaison méiotique a pu être démontré chez les organismes modèles, comme la souris. *RNF212* est localisé dans une région centrale du complexe synaptonémal et joue un rôle clé dans la réparation des *DSBs* et notamment leur résolution en crossing-over de type I (Lake et Hawley, 2013). En effet, chez des souris mutées pour ce gène, il n'y a pas de formation des complexes spécifiques à la constitution des crossing-overs (Qiao *et al.*, 2014). Au stade Pachytène, il se fixe sur le complexe synaptonémal afin de stabiliser le complexe *MSH4-MSH5* (MutSγ), impliqué dans la formation des doubles jonctions de Holliday. L'action antagoniste d'un autre gène, *HEI10*, est nécessaire pour permettre le désassemblage de *RNF212* et de *MSH4* et *MSH5* du complexe synaptonémal et la bonne formation des crossing-overs (Qiao *et al.*, 2014). De plus, *RNF212* pourrait jouer un rôle de rétro-contrôle ; la fixation du complexe MutSγ, dépendante de *RNF212*, pourrait augmenter, à son tour, la fixation de *RNF212* (Reynolds *et al.*, 2013).

RNF212 est très proche d'un autre gène, cité précédemment, *CPLX1*, cela pourrait expliquer pourquoi ce dernier est statistiquement associé à la variation inter-individuelle du taux de recombinaison, alors qu'aucune fonction liée à la recombinaison n'est connue pour ce gène (Kong *et al.*, 2014). Il se peut donc que ce soit un faux positif.

IV. 3. *Le déterminisme génétique de la variation inter-individuelle de la localisation de la recombinaison*

A ce jour, il semblerait qu'il n'y ait qu'un seul gène identifié influençant la localisation et

l'utilisation des points chauds. Il s'agit de *PRDM9*, préalablement cité, car à l'origine de la formation des *DSBs* chez certains Mammifères. Ce gène est exprimé spécifiquement dans les méiocytes, c'est-à-dire les cellules s'engageant dans le processus de la méiose. En plus de son domaine codant pour la formation de doigts de zinc, il possède également une activité méthyltransférase qui est à l'origine de la triméthylation de la lysine 4 de la protéine histone *H3* (*H3K4me3*) (Baudat *et al.*, 2013). Chez la souris, la majorité des points chauds de recombinaison est déterminée par *PRDM9*. Plusieurs faits prouvent que ce gène est impliqué dans la détermination des points chauds de recombinaison. Tout d'abord, les points chauds actifs chez la souris sont enrichis en *H3k4me3* dans les spermatocytes, ensuite les doigts de zinc de la protéine se lient *in vitro* avec des motifs présents au centre des points chauds (Baudat *et al.*, 2013), enfin des polymorphismes au sein du gène influent sur l'activité des points chauds entre les individus, chez la souris (Parvanov *et al.*, 2010) et chez l'Homme (Berg *et al.*, 2010). En effet, chez l'Homme, *PRDM9* présente plusieurs allèles conduisant à un nombre différent de doigts de zinc (entre 8 et 18). La plus grande diversité est observée au sein de la population africaine, par rapport à la population européenne où un allèle est présent à plus de 80 %. Des hommes portants des allèles différents montrent des fréquences de crossing-overs différentes, allant jusqu'à une réduction de 5 % de l'activité normale des points chauds (Berg *et al.*, 2010). Une première hypothèse sur l'incapacité de certains allèles à activer des points chauds reposerait sur la constitution des doigts de zinc, incapables de reconnaître des motifs connus. Or, plusieurs allèles présentent des doigts de zinc très semblables à l'allèle activateur. Les allèles de *PRDM9* sont également capables d'influencer différents points chauds (Berg *et al.*, 2010). Il semblerait donc que les variations au sein des doigts de zinc de *PRDM9* influencent fortement l'activité des points chauds.

Cependant, *PRDM9* n'est pas uniquement le seul gène connu pour la détermination des points chauds, c'est également le seul « gène de spéciation » connu chez les Vertébrés. En effet, il a été montré que l'hybride résultant de l'accouplement de deux sous-espèces de souris était stérile. Cette stérilité serait due à des problèmes d'appariement des chromosomes homologues et donc à une méiose stoppée (Davies *et al.*, 2016). *PRDM9* est à l'origine de la formation des *DSBs*, or il y a beaucoup d'évènements mutationnels au niveau des doigts de zinc qui changent complètement ses sites de fixation et donc la localisation des points chauds et des *DSBs*. Ces modifications de la localisation des points chauds rendent impossible l'appariement des

chromosomes entre deux sous-espèces (Davies *et al.*, 2016). En revanche, l'ajout du gène *PRDM9* humain permet de restaurer la fécondité. Etant donné que l'allèle humain n'est pas présent chez la souris, les sites de fixation n'ont pas connu de modifications, d'érosions et donc les *DSBs* sont majoritairement symétriques entre les sous-espèces.

Malgré son importance, le gène *PRDM9* n'est pas présent chez toutes les espèces. Il est effectivement absent chez les plantes (Edlinger et Schlögelhofer, 2011), chez les oiseaux (Pontig, 2011) et même chez certains Mammifères ; les canidés (Auton *et al.*, 2013). Chez le chien, il existe bien un orthologue de *PRDM9*, mais il a subi plusieurs mutations qui ont tronqué le dernier exon[9], codant pour les doigts de zinc, ce qui le rend non fonctionnel (Campbell *et al.*, 2016). Dans cette espèce, comme pour les souris mutantes pour *PRDM9*, la recombinaison tend à se regrouper au niveau des promoteurs des gènes et des îlots *CpG* (ce qui se retrouve également chez *Arabidopsis* ou l'oiseau (Campbell *et al.*, 2016)). En revanche, les patrons de recombinaison restent similaires aux autres espèces de Mammifères avec la majorité des crossing-overs au niveau des régions télomériques. Cependant, contrairement à l'Homme, la répartition de la recombinaison chez le chien est plus homogène sur le génome, indiquant que *PRDM9* concentrerait la recombinaison au niveau de zones spécifiques, les points chauds (Campbell *et al.*, 2016).

Le gène *PRDM9* a également été détecté chez la vache. Dans une première étude, il a été supposé être présent sur le chromosome X (Sandor *et al.*, 2012), cependant une étude secondaire a permis de montrer qu'il s'agissait en fait probablement d'un paralogue et que le gène *PRDM9* était effectivement sur le chromosome 1 (Ma *et al.*, 2015).

Dans l'étude de Ma et collaborateurs (2015), *PRDM9* semblait plutôt avoir un rôle sur la variation inter-individuelle du taux de recombinaison, mais un tel résultat n'a pu être retrouvé pour d'autres espèces d'élevage. Dans le cas de l'expérience de Sandor et collaborateurs (2012), il a effectivement pu être lié à un phénotype de biais d'usage des points chauds. Cependant, l'étude était assez peu résolutive et le résultat est donc à interpréter avec précaution.

---

9 Les exons sont des portions d'un gène transcrites en *ARN* et qui sont conservées dans l'*ARN* messager après épissage.

IV. 4. *Etat de l'art chez le mouton*

Jusqu'à très récemment, la recombinaison génétique chez le mouton a été assez peu étudiée. Une première carte génétique a été proposée en 2007. Elle mesurait environ 3 580 cM et comprenait 1 374 marqueurs représentant 1 333 loci (Maddox et Cockett, 2007). Cependant, les deux phénotypes de recombinaison évoqués n'avaient pas encore été étudiés. Il faudra attendre 2016 pour l'obtention des premiers résultats. En effet, le phénotype de variation inter-individuelle du taux de recombinaison a alors été étudié dans la race de mouton sauvage, le Soay (Johnston *et al.*, 2016). Cette population fait partie du groupe des « races du Nord » et est issue de la première vague de domestication du mouton, aux environs de 11 000 ans avant J. C.

L'étude de la variation de la recombinaison chez les Soay a permis de préciser la longueur de la carte génétique qui est de 3 304 cM pour les deux sexes, et a permis de montrer que le mouton partageait également des patrons de recombinaison communs aux autres animaux précédemment évoqués (Johnston *et al.*, 2016). Le taux de recombinaison dépend là-aussi de la taille des chromosomes : plus les chromosomes sont petits, plus le taux de recombinaison est important. De plus, les plus forts taux de recombinaison sont observés à proximité des télomères et ils diminuent lorsque l'on s'en éloigne. L'étude a exclu le potentiel effet de facteurs environnementaux, tels que l'âge des parents ou le coefficient de consanguinité sur le taux de recombinaison, mais a montré qu'il y avait bien une variation entre les individus. Cette variation est également héritable, avec une valeur d'héritabilité de 0,15, ce qui est similaire à celle de la vache (Sandor *et al.*, 2012).

L'équipe de Johnston a également cherché à connaître le déterminisme génétique de ce phénotype. Ils ont réalisé une première *GWAS* qui a permis d'identifier différents loci. Notamment un majeur sur le chromosome 6 (ensuite précisé par imputation) et qui correspond à une région comportant plusieurs gènes candidats, tels que *RNF212*, *CPLX1*, *GAK* et *PCGF3*. Suite à cela, ils ont augmenté leur résolution en utilisant une méthode haplotypique, ce qui leur a permis de détecter un deuxième *QTL* sur le chromosome 7, à proximité des gènes *RNF212B* et *REC8*, préalablement cités.

Il n'y a pas encore eu d'études sur les points chauds, leur utilisation ou leur déterminisme génétique chez le mouton.

V. Objectifs de la thèse

Au début de la thèse, la recombinaison chez les ovins n'avait été que peu étudiée et son déterminisme génétique était encore inconnu. Nous disposions de nombreuses données pour étudier la recombinaison génétique et son déterminisme génétique, notamment en race Lacaune, du fait de la mise en place de la sélection génomique[10] en 2015, ainsi que grâce à un projet de détection de *QTL* dans de grandes familles. On avait ainsi accès à plus de 6 000 génotypages moyenne densité (environ 50 000 marqueurs) en race Lacaune. La mise en place d'un projet d'étude de diversité génétique a également permis d'obtenir des génotypages haute densité (autour de 600 000 marqueurs) pour une soixantaine d'autres moutons Lacaune.

La mise à disposition de ces différentes données a rendu possible ce projet de thèse d'étude de la recombinaison génétique, d'un point de vue familial, avec la recherche d'une potentielle variation du taux de recombinaison, mais également d'un point de vue populationnel. Cette dernière possibilité permettait la recherche d'éventuels points chauds et d'un possible biais dans leur utilisation par les différents individus.

Le but de cette thèse était donc ***l'étude des zones de recombinaison du génome, de leur déterminisme génétique et des intérêts en sélection génomique chez les ovins.***

Pour cela, le travail a été organisé en trois parties :

Première partie

Au cours de cette première partie, la distribution de la recombinaison sur le génome a été étudiée : observation de la répartition des crossing-overs, recherche de points chauds et étude du biais d'usage des points chauds. Ce phénotype a été étudié de deux manières différentes, une première approche directe où nous avons cherché à estimer l'utilisation pour chaque individu des éventuels points chauds détectés et une deuxième approche, un peu plus indirecte où nous avons

---

10 Sélection basée sur une estimation de valeurs génétiques des candidats à partir de l'information donnée par les marqueurs denses couvrant tout le génome.

recherché de potentiels motifs d'*ADN* qui pourraient être reconnus par *PRDM9*. Cette deuxième approche a fait le cadre d'un stage de M1 en bio-informatique.

Deuxième partie

Le phénotype du taux de recombinaison individuel a été étudié dans un second temps. Pour cela, il a tout d'abord fallu estimer un taux de recombinaison individuel après avoir détecté des crossing-overs au sein de familles. Des variations ayant été observées entre ces taux de recombinaison individuels, il a donc été possible de rechercher des loci pouvant les expliquer. Pour cela, il a d'abord fallu utiliser l'imputation pour augmenter la densité des génotypages des moutons issus des données familiales, puis une *GWAS* pour rechercher de potentiels loci ayant un effet significatif sur le phénotype.

Troisième partie

Cette dernière partie s'est plus axée sur l'utilisation qui pouvait être faite de la recombinaison génétique en sélection. Les distances génétiques entre marqueurs ont été utilisées pour créer des ensembles de *SNPs* de faible densité et voir si l'utilisation de cette information pouvait rendre l'imputation sur la puce à *SNPs* de moyenne densité, la puce 50K, plus efficace que celle réalisée avec les puces basse densité existantes. Un tel résultat permettrait de concevoir des puces de très basse densité, donc moins chères et plus intéressantes pour les éleveurs.

Chapitre 2 : Création de cartes de recombinaison

Chapitre 2 : Utilisation de deux jeux de données différents pour créer des cartes génétiques de haute résolution

I. Deux jeux de données disponibles en Lacaune

I. 1. *La race ovine Lacaune lait : sujet de l'étude*

I. 1. a. *Généralités sur la race ovine Lacaune lait*

La race ovine Lacaune tire son nom du chef-lieu du canton situé au milieu des monts de Lacaune dans le Tarn. Ce département, avec l'Aveyron et quelques départements limitrophes, constituent le berceau de la race, le « Rayon de Roquefort ». En effet, son lait est essentiellement utilisé pour la production de fromage Roquefort.

La race Lacaune est aujourd'hui la première race ovine française en termes d'effectifs (presque 60% du cheptel ovin au contrôle laitier). En effet 930 000 brebis ont été traites en 2012 dont 171 909 au *CLO*[11] (Contrôle Laitier Officiel) et 490 685 au *CLS*[12] (Contrôle Laitier Simplifié) en 2013 *(source : Résultats de Contrôle Laitier de l'Idele, 2013).* Les brebis sont réparties dans 367 troupeaux. La race est gérée par un Organisme de Sélection[13] *(OS)* et par deux Entreprises de

---

11 Le contrôle laitier officiel est un contrôle qui est exclusivement réalisé chez les éleveurs sélectionneurs. Il inclut un contrôle quantitatif (quantité de lait) et un contrôle qualitatif (taux de matière grasse et de protéines).

12 Le contrôle laitier simplifié est un contrôle effectué dans les élevages utilisateurs du progrès génétique. Il ne concerne que la quantité de lait.

13 Un organisme de sélection est une association chargée d'organiser la sélection d'une race animale. Il regroupe les différents acteurs concernés par cette race. L'organisme de sélection

Sélection[14] *(ES)* : Ovitest et le Service Elevage de la Confédération Générale de Roquefort. Bien qu'elle présente le plus grand nombre d'individus, l'élevage de la race Lacaune reste local, relativement confiné dans son berceau. Cependant, elle s'exporte quand même dans de nombreux pays : Portugal, Espagne, Grèce, Allemagne etc.

Dès le début du XXe siècle, la race commence à être sélectionnée pour sa production laitière, qui était alors de quelques dizaines de litres par an, pour 288,5L en 167 jours aujourd'hui *(source : Résultats de Contrôle Laitier de l'Idele, 2013).* Puis après la Seconde Guerre Mondiale, le processus de sélection s'est accéléré avec plusieurs étapes clés. Il a tout d'abord porté sur la productivité des brebis et des troupeaux avec, comme critères de sélection, l'augmentation de la production laitière à la traite ainsi que des quantités de matière grasse et protéique, donc l'augmentation des quantités de fromage. Une telle sélection, lorsqu'elle devient efficace (dans les années 80), entraine une dégradation de la richesse du lait au niveau des taux butyreux et protéique. Les critères de sélection évoluent donc vers une amélioration de la composition chimique du lait et de ses aptitudes fromagères (décennies 80 et 90). Puis lorsque la productivité des femelles n'était plus la priorité, l'intérêt s'est porté dans les années 2000 sur les caractères fonctionnels, tels l'aptitude à la traite mécanique (morphologie de la mamelle) ou la résistance génétique aux maladies comme les mammites. Ceci afin de réduire les coûts de production et d'améliorer les conditions de travail, notamment lors de la traite. Aujourd'hui, la sélection génomique est de mise. Elle a démarré au début des années 2000 par la sélection d'un gène à effet majeur sur la résistance à la tremblante, via le génotypage à grande échelle du gène *PrP*. Depuis 2003, la totalité des béliers qualifiés est homozygote résistante (R/R). La sélection génomique s'est poursuivie avec la mise en place d'un schéma génomique lors de la campagne 2015, à la suite du programme ROQUEFORT'IN.

---

définit les caractéristiques de la race, gère son livre généalogique et fixe ses objectifs de sélection.

14 Une entreprise de sélection gère la production et la diffusion des paillettes des reproducteurs mâles d'insémination animale d'une race issus du programme d'amélioration génétique de cette race.

I. 1. b. *La mise en place de la sélection génomique en Lacaune*

Depuis 2014-2015, la race Lacaune bénéficie d'un schéma de sélection génomique à la suite du projet ROQUEFORT'IN. Elle permet d'accélérer le progrès génétique en prédisant précocement la valeur génétique des individus à partir de leur *ADN*. Un des produits secondaires de cette sélection est la possibilité d'utiliser les phénotypes et les génotypes pour l'identification des gènes d'intérêt (Sallé et Moreno, 2011, Rupp *et al.*, 2016). L'opportunité de mettre en place une sélection génomique en ovin date de 2009 avec le développement par le consortium international de génomique ovine d'une puce génomique de 54 241 marqueurs, commercialisée par Illumina. Les différents effets des marqueurs sont combinés dans des équations de prédiction établies à partir d'une population de référence suffisamment grande. Ces équations permettront de connaître la valeur génétique des individus à sélectionner, sans connaître leurs performances, juste leur génotype.

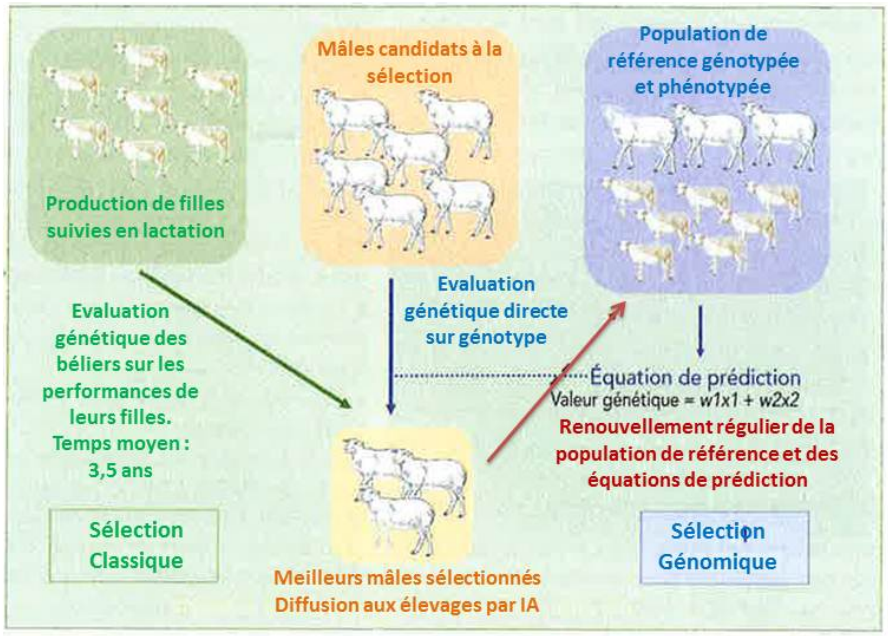

Figure 7 : Comparaison entre la sélection classique et la sélection génomique (d'après Moreno et Sallé, 2011)

La sélection génomique conduit à une indexation directe des béliers à partir de leur génotype et des équations de prédiction. Ceci se traduit par un gain de temps par rapport au schéma de sélection classique (**voir Figure 7**). Notamment grâce à une réduction de l'intervalle de

génération pour obtenir des prédictions fiables pour des caractères comme les paramètres de lactation. Grâce à la sélection génomique, une valeur génétique est acquise dès la naissance du bélier et est équivalente à celle obtenue si le bélier avait produit 30 à 150 filles testées dans le schéma classique (Sallé et Moreno, 2011). En sélection classique, l'intensité de sélection pour la principale voie de sélection (père-fils) s'exerce relativement tardivement : vers 2 ans et demi, à l'issue des résultats de testage sur descendance. Cela demande l'entretien de plusieurs béliers en attente d'index, alors qu'un petit nombre sera réellement utilisé. La sélection génomique permet d'exercer une pression de sélection plus tôt, dès que les index génomiques sont disponibles (vers 3 mois). Il existe cependant une limite à la sélection génomique, c'est la taille de la population de référence. En effet, il est nécessaire d'avoir un minimum d'individus génotypés (généralement aux environs de 1 000 individus, cependant, cela dépend de la taille efficace de la population) pour avoir des valeurs génétiques fiables et cette population doit être renouvelée régulièrement. Cet effectif est atteint en race Lacaune avec, en 2009, 2 840 béliers testés et génotypés (Astruc *et al.*, 2012).

En race Lacaune, l'*OS* a fait le choix d'une sélection à coûts constants avec un progrès génétique annuel légèrement supérieur à celui obtenu par la sélection classique : accroissement d'environ 20% (Astruc *et al.*, 2012).

La mise en place de la sélection génomique en race Lacaune a permis d'obtenir un grand nombre de données de génotypages, notamment grâce à l'existence de la puce à *SNPs* moyenne densité avec 54 000 marqueurs (*SNPs*). De plus, un deuxième projet d'étude de la diversité génétique chez les moutons, l'Action Innovante de France Génétique Elevage « Obtention des Parentés par Assignation » (*OPA*) a également permis d'obtenir une deuxième puce, cette fois-ci de haute densité, car elle contient plus de 600 000 *SNPs* (Tortereau *et al.*, 2014).

I. 2. *Un grand pédigrée génotypé avec une puce moyenne densité*

L'existence de ces deux puces à *SNPs* a permis d'exploiter deux jeux de données différents et indépendants au cours de cette thèse, notamment un en données familiales. Il est composé de 8085 animaux apparentés qui ont tous été génotypés avec la puce moyenne densité : Illumina Ovine Beadchip® comprenant 54 241 *SNPs* (puce 50K).

Suite à plusieurs étapes de contrôle qualité au cours desquelles nous avons exclu les animaux ayant un « call rate » inférieur à 95% et ne respectant pas le principe de la ségrégation mendélienne, nous avons conservé 5 940 individus ayant leur père connu et génotypé pour 46 813 marqueurs. Ces marqueurs ont été obtenus après avoir exclu les *SNPs* avec un « call freq » inférieur à 98%. Les 5 940 animaux descendaient de 345 pères et étaient divisés selon deux pédigrées (**voir Figure 8**). Ces pères ont été choisis selon deux critères : soit ils possédaient au moins 2 descendants génotypés et leur propre père était génotypé, soit leur père n'était pas génotypé, mais ils avaient au moins 4 descendants génotypés. Etant donné qu'il n'y a que des mâles parmi les 345 pères, ce ne sont que des méioses mâles qui seront étudiées.

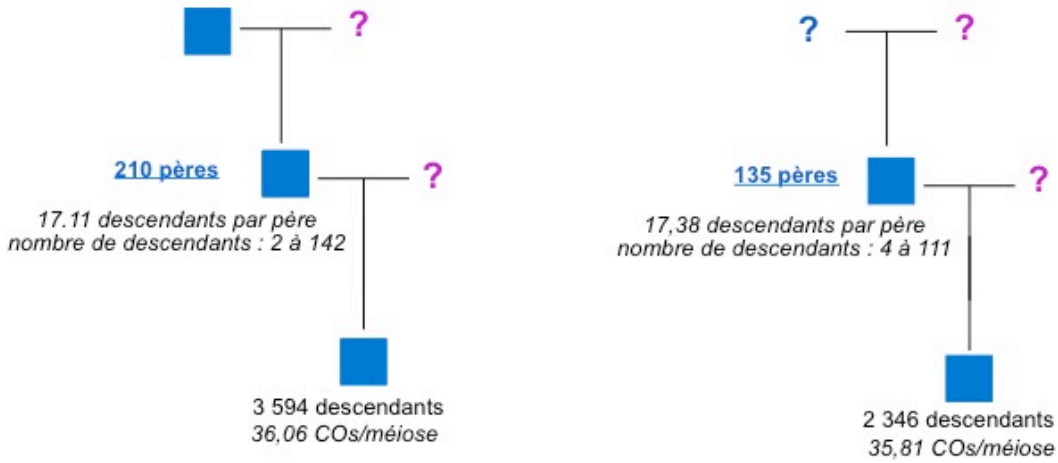

Figure 8 : Répartition des 5 940 Lacaune dans deux pédigrées

Dans les deux cas, les animaux n'ont pas leur mère de génotypée. En revanche, dans le pédigrée de gauche, les descendants ont leur grand-père de génotypé. Les pères ont tous au moins 2 descendants de génotypés.

I. 3. *Un échantillon d'animaux non apparentés génotypés avec une puce haute densité*

Le deuxième jeu de données disponible pour la thèse était issu du projet *OPA* visant à réaliser des travaux sur l'assignation de parenté en races ovines (laitières et bouchères) françaises. Il était notamment composé de 70 béliers Lacaune non apparentés et choisis pour représenter au mieux la race Lacaune et la diversité génétique de cette population. Les deux rameaux Lacaune ; Lacaune viande et Lacaune lait, finalement très proches, ont été utilisés (Rochus *et al.*, 2017). Les

animaux ont tous été génotypés avec la puce haute densité Illumina Ovine® Infinium HD SNP comprenant 685 734 *SNPs* (Moreno-Romieux *et al.*, 2017) (puce 600K). Un test pour l'Equilibre de Hardy Weinberg a été effectué pour chaque *SNP* et ceux qui ne respectaient pas cet équilibre étaient supprimés. De même que les marqueurs avec une *MAF* (Minor Allele Frequency) à 0 ou un « call rate » inférieur à 99%. Suite aux étapes de contrôle qualité, 503 784 *SNPs* ont été conservés.

Nous avons comparé l'apparentement entre les animaux du pédigrée et les 70 Lacaune non apparentés. Trente-deux animaux étaient non apparentés (coefficient de parenté < 0,2), 3 animaux étaient cousins (avec un coefficient d'apparentement compris entre 0,2 et 0,3) et 16 animaux étaient plutôt fortement apparentés (probablement une relation père-descendant) puisqu'ils avaient des coefficients d'apparentement entre 0,45 et 0,6. Les 19 animaux restants étaient en commun entre les 70 Lacaune et le fichier pédigrée. Nous les avons donc conservés dans le pédigrée, mais supprimés de l'échantillon d'animaux non apparentés pour construire les cartes populationnelles. Ce qui fait que nous avons réalisé les analyses des cartes populationnelles avec seulement 51 animaux.

II. L'étude des cartes de recombinaison en données familiales

II. 1. *La détection des crossing-overs*

Nous avons utilisé le logiciel LINKPHASE pour réaliser la détection des crossing-overs. Le principe de ce logiciel a été expliqué dans le Chapitre 1, dans le paragraphe « II. 1. a. b. ». Brièvement, il utilise des informations familiales (des parents génotypés ou plusieurs descendants génotypés) afin de reconstruire des haplotypes en utilisant les lois de ségrégation mendélienne et les informations de parenté. Pour combler de potentielles lacunes, il utilise ensuite les modèles de Markov cachés (Druet et Georges, 2010). Après avoir fait tourner une première fois LINKPHASE, nous avons supprimé les « doubles crossing-overs », c'est-à-dire des crossing-overs qui ont lieu dans la même méiose à moins de 3 Mb l'un de l'autre. Nous avons choisi cet intervalle car il correspondait clairement à une aberration dans la distribution des distances entre crossing-overs. Nous avons considéré que les génotypages compris dans cet intervalle de 3 Mb étaient manquants

et nous avons refait tourner LINKPHASE pour estimer de nouveaux taux de recombinaison.

Les fichiers de sortie de LINKPHASE permettent d'obtenir le nombre de crossing-overs entre un père et ses descendants, et donc d'obtenir un nombre de crossing-overs total pour chaque père. Cela permet donc de connaître le nombre de crossing-overs obtenus dans chacun des deux pédigrées (**voir Figure 9**).

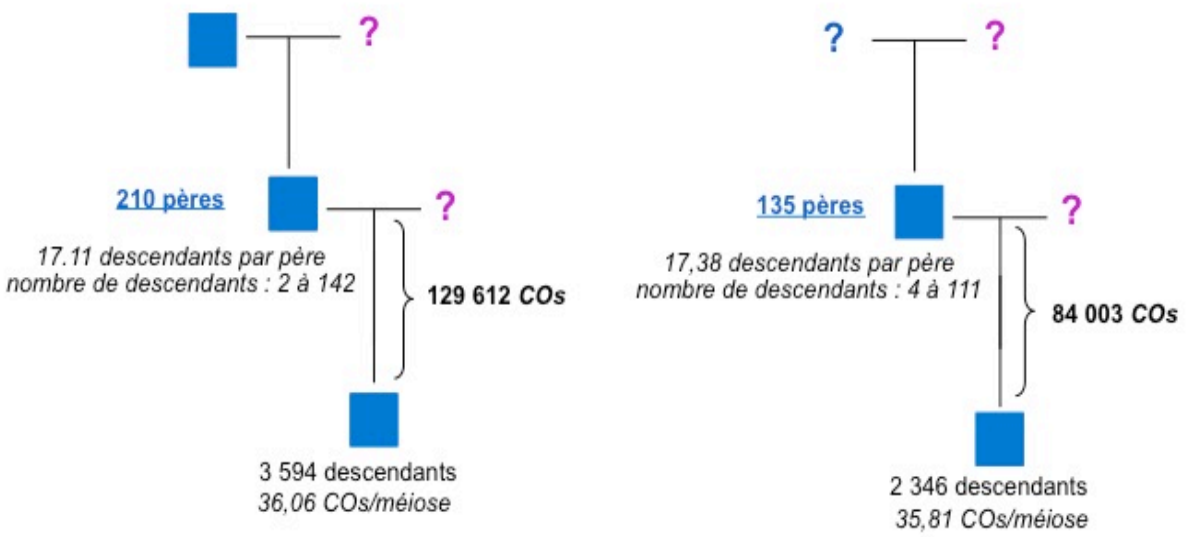

Figure 9 : Nombre de crossing-overs obtenus dans les deux pédigrées

Les crossing-overs ont été identifiés grâce aux méioses de 345 pères. Deux-cent-dix avaient leur père de génotypé, tandis que les 135 autres ne l'avaient pas.

Il a ainsi été possible d'identifier 213 615 crossing-overs au total. Bien que 135 pères n'aient pas eux-mêmes leur propre père de génotypé, cela n'a pas affecté notre capacité à détecter les crossing-overs, en effet le nombre moyen de crossing-overs par méiose est similaire dans les deux pédigrées (36,1 pour ceux qui ont leur père de génotypé, 35,8 pour les autres). De plus, le nombre de descendants n'a pas d'effet statistique sur le nombre moyen de crossing-overs par méiose ($p > 0,23$). Ceci peut s'expliquer par le fait que les 135 pères qui n'ont pas leur père de génotypé possèdent un nombre de descendants plus important (17,4 en moyenne et le nombre de descendants allant de 4 à 111 ; **voir Figure 8**). Ce nombre important de descendants permet de reconstruire correctement les haplotypes des pères.

Suite à l'identification des crossing-overs, il a ensuite été possible de définir un phénotype

individuel qui correspond au nombre moyen de crossing-overs pour un père sur l'ensemble de ses méioses (*i. e.* de ses descendants).

II. 2. *L'estimation des taux de recombinaison*

Une fois que les crossing-overs sont connus, il est possible d'estimer des taux de recombinaison familiaux dans des fenêtres de 1 Mb et dans les intervalles entre les marqueurs de la puce 50K grâce à une méthode statistique inspirée de Cheung *et al.* (2007). Pour les intervalles génétiques de petite taille, comme ceux considérés ici, le taux de recombinaison (appelé ***c*** dans les équations suivantes) est souvent exprimé en centimorgans par mégabase (0,01 Morgan par mégabase). Ainsi, la probabilité, mesurée en Morgan, qu'un crossing-over ait lieu dans une méiose et dans un intervalle ***j*** (ici les intervalles de la 50K ou des intervalles de 1 Mb) est :

$$\mathbf{0{,}01 * c_j * l_j} \text{ avec :}$$

- $\mathbf{c_j}$: taux de recombinaison dans l'intervalle ***j***.
- $\boldsymbol{l_j}$: longueur de l'intervalle ***j*** exprimée en mégabase.

Lorsque les ***M*** méioses d'un individu sont considérées, le nombre attendu de crossing-overs est alors de :

$$\mathbf{0{,}01 * c_j * l_j * M}$$

Pour estimer le taux global de recombinaison de la population, il faut pouvoir combiner les nombres attendus de crossing-overs de tous les individus. Cependant, lorsque cette analyse est faite, il faut tenir compte du fait que les individus ***s*** ont un nombre moyen de crossing-overs par méiose différent, appelé $\boldsymbol{R_s}$ qui peut biaiser le taux de recombinaison global, surtout si ce nombre moyen de crossing-overs est extrême. Pour compenser ce potentiel biais, il faut exprimer $\boldsymbol{R_s}$ comme une proportion du nombre moyen total de crossing-overs par méiose sur l'ensemble des individus, ***R***, ce qui revient à calculer $\boldsymbol{R_s/R}$. Et donc, finalement, pour un individu ***s*** et un intervalle ***j***, le nombre attendu de crossing-overs est égal à :

$$\mathbf{0{,}01 * c_j * l_j * M * (R_s/R)}$$

Dans la littérature, une façon naturelle d'estimer la distribution des crossing-overs est d'utiliser une loi de Poisson. En effet, elle est principalement utilisée pour décrire le

comportement d'un certain nombre d'évènements se produisant notamment dans un intervalle donné, avec une fréquence moyenne connue et indépendamment des évènements ayant eu lieu dans l'intervalle précédent. C'est donc une bonne manière d'exprimer la distribution des crossing-overs si nous estimons qu'ils sont indépendants les uns des autres et donc non soumis au phénomène d'interférence.

En estimant que le nombre y_{sj} de crossing-overs observés dans un intervalle j pour un individu s suit une loi de Poisson, alors son paramètre λ, c'est-à-dire le nombre moyen de crossing-overs dans l'intervalle, est donné par :

$$y_{sj} \mid c_j \sim \text{Poisson}[(0{,}01 * l_j * c_j * M_s * (R_s/R)] \quad (1)$$

Avec cette paramétrisation, c_j est exprimé en cM/Mb et la probabilité $P(y_{sj} \mid c_j)$ est une vraisemblance. Lorsque nous voulons combiner tous les crossing-overs parmi tous les individus, la vraisemblance de c_j correspond au produit des vraisemblances de Poisson, estimées pour chaque individu et issues de l'équation **(1)**. Pour estimer les taux de recombinaison, il est possible d'utiliser un *a priori*. En effet, l'utilisation d'un *a priori* permet de considérer que le taux de recombinaison n'est pas identique sur le génome et qu'il existe des régions où la recombinaison est supérieure (ou inférieure) à la moyenne globale.

Ainsi, la distribution *a priori* de c_j est définie comme suit :

$$c_j \sim \Gamma(\alpha, \beta) \quad (2)$$

Pour estimer α et β, nous calculons les c_j approchés en utilisant la méthode de Sandor *et al.* (2012) sur tout le génome, puis nous ajustons une distribution gamma sur la distribution observée, ce qui permet d'obtenir la distribution *a priori* des c_j. En combinant l'*a priori* **(2)** avec les vraisemblances de l'équation **(1)**, on obtient les distributions *a posteriori* de c_j :

$$c_j \mid y_j \sim \Gamma[\alpha + \sum_s y_{sj},\ \beta + \sum_s 0{,}01 * l_j * M_s * (R_s/R)] \quad (3)$$

Ces distributions *a posteriori* permettent d'estimer le taux de recombinaison. Nous pouvons noter que la moyenne *a posteriori* de c_j s'écrit **(α + nb *COs*) / (β + nb méioses)**. En effet, le terme $\Sigma(y)$ correspond au nombre de crossing-overs observés dans l'intervalle et le terme $\Sigma[Ms*(Rs/R)]$ correspond au nombre de méioses incluant une pondération pour les variations inter-individuelles du taux de recombinaison. Ceci montre l'influence du *prior* sur notre calcul : lorsque le nombre de crossing-overs et le nombre de méioses sont faibles, cette moyenne *a posteriori* tend vers α/β,

c'est-à-dire la moyenne *a priori*. Inversement quand ces deux nombres sont élevés les termes α et β perdent de leur importance et le taux de recombinaison estimé tend vers l'estimateur empirique ***(nb COs / nb méioses)***.

Puisque la localisation des crossing-overs n'est souvent pas suffisamment précise pour pouvoir les assigner avec certitude à un unique intervalle génomique, les estimations finales de c_j sont obtenues après trois étapes :

- 1) pour chaque crossing-over chevauchant l'intervalle j et localisé au sein d'une fenêtre de longueur L, nous définissons x_c comme une variable prenant la valeur de 1 si le crossing-over a lieu dans l'intervalle j et 0 sinon. Nous supposons que, localement, le taux de recombinaison est proportionnel à la distance physique et nous posons :

$$P(x_c = 1) = \min(l_j/L, 1)$$

- 2) en utilisant la probabilité établie au stade **(1)**, nous échantillonnons x_c pour chaque crossing-over chevauchant l'intervalle j, parmi toutes les possibilités possibles pour un crossing-over, et nous posons :

$$y_{sj} = \sum_c x_c$$

- 3) Sachant y_{sj}, nous échantillonnons les c_j à partir de l'équation **(3)**.

Pour chaque intervalle considéré, nous réalisons les étapes **(2)** et **(3)** environ 1 000 fois afin d'obtenir les échantillons de la distribution *a posteriori* de c_j, prenant ainsi en compte l'incertitude due au manque de précision de la localisation des crossing-overs.

Cette méthode a permis d'estimer un taux moyen de recombinaison sur le génome d'environ 1,5 cM/Mb, sachant que la taille du génome couverte par la puce 50K est de 2,45 Gb.

II. 3. *Les Lacaune ont des patrons de recombinaison communs aux animaux d'élevage*

Les taux de recombinaison ont été estimés pour des intervalles de 1 Mb, ainsi que pour les intervalles entre les marqueurs de la puce 50K. Dans les deux cas le taux de recombinaison mesuré sur une région spécifique d'un chromosome dépend principalement de sa position relative au télomère. En effet, les régions proches des télomères sont celles qui recombinent le plus, sachant

qu'aux télomères même, la recombinaison est plus faible. C'est encore plus vrai, au niveau des centromères, la recombinaison est beaucoup plus faible (**voir Figure 10**). Ces caractéristiques sont vraies pour les chromosomes métacentriques[15] (les trois premiers chez le mouton), mais aussi pour les chromosomes acrocentriques[16].

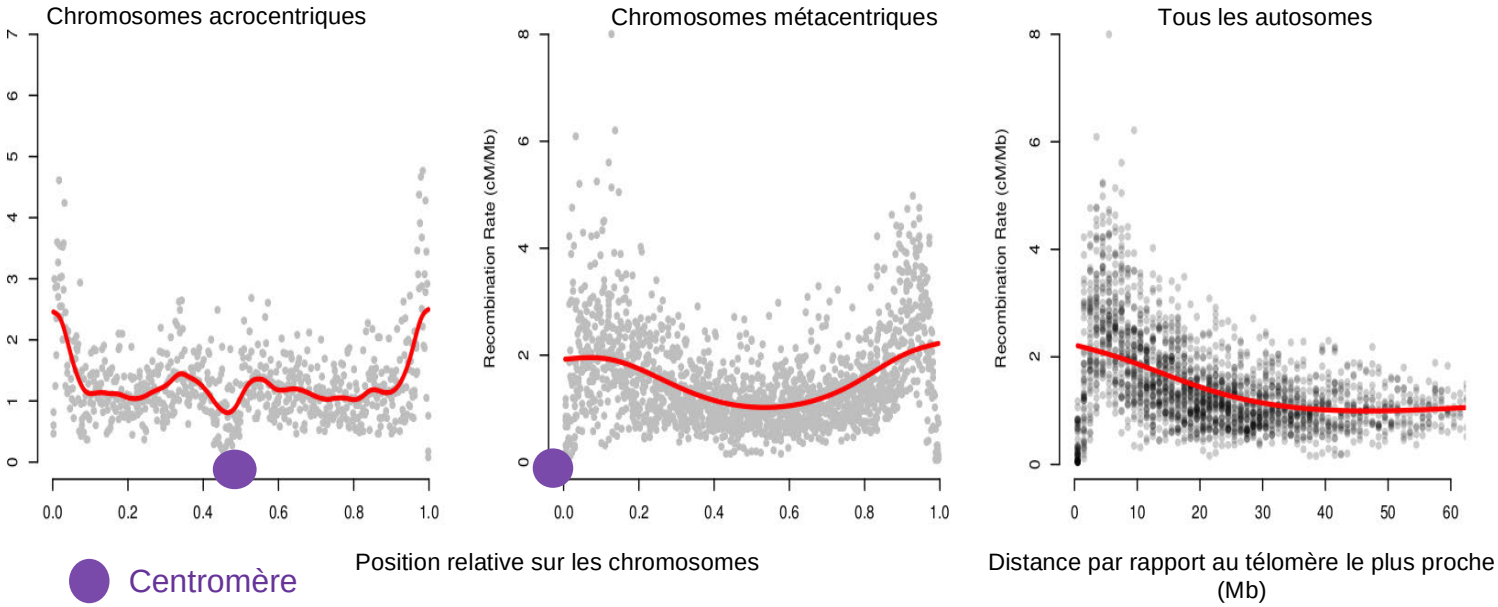

Figure 10 : Patrons de recombinaison le long des autosomes des Lacaune

Graphe de gauche : taux de recombinaison sur des fenêtres de 1 Mb le long des chromosomes métacentriques (chromosomes 1, 2 et 3). Au centre : taux de recombinaison sur des fenêtres de 1 Mb le long des chromosomes acrocentriques (4 à 26). A droite : taux de recombinaison sur des fenêtres de 1 Mb en fonction de la distance au télomère le plus proche. Les deux premiers graphiques ont été obtenus en réajustant les échelles afin de pouvoir les comparer et les cumuler en un chromosome unique.

Nous avons également pu observer que la recombinaison restait faible dans les intervalles compris dans les 4 Mb les plus proches des fins de chromosomes. Cette diminution pourrait notamment être due à un manque de crossing-overs non détectés au niveau des extrémités, du

---

15 Chromosome pour lequel le centromère divise la chromatide en deux bras égaux, il est en position médiane.

16 Chromosome dont le centromère est situé près d'une extrémité : souvent un seul bras chromosomique.

fait d'un manque d'informativité des marqueurs au niveau des télomères. Dans les analyses suivantes, nous supprimerons donc ces régions afin d'éviter de potentiels biais.

Pour les deux types d'intervalle considérés ; fenêtres de 1 Mb et intervalles de la 50K, nous avons estimé les taux de recombinaison spécifiques pour chaque chromosome (**voir Figure 11**). On remarque que, comme pour la majorité des Mammifères, les Lacaune ont un taux de recombinaison qui dépend de la taille des chromosomes ; plus les chromosomes sont petits, plus leur taux de recombinaison augmente, et inversement.

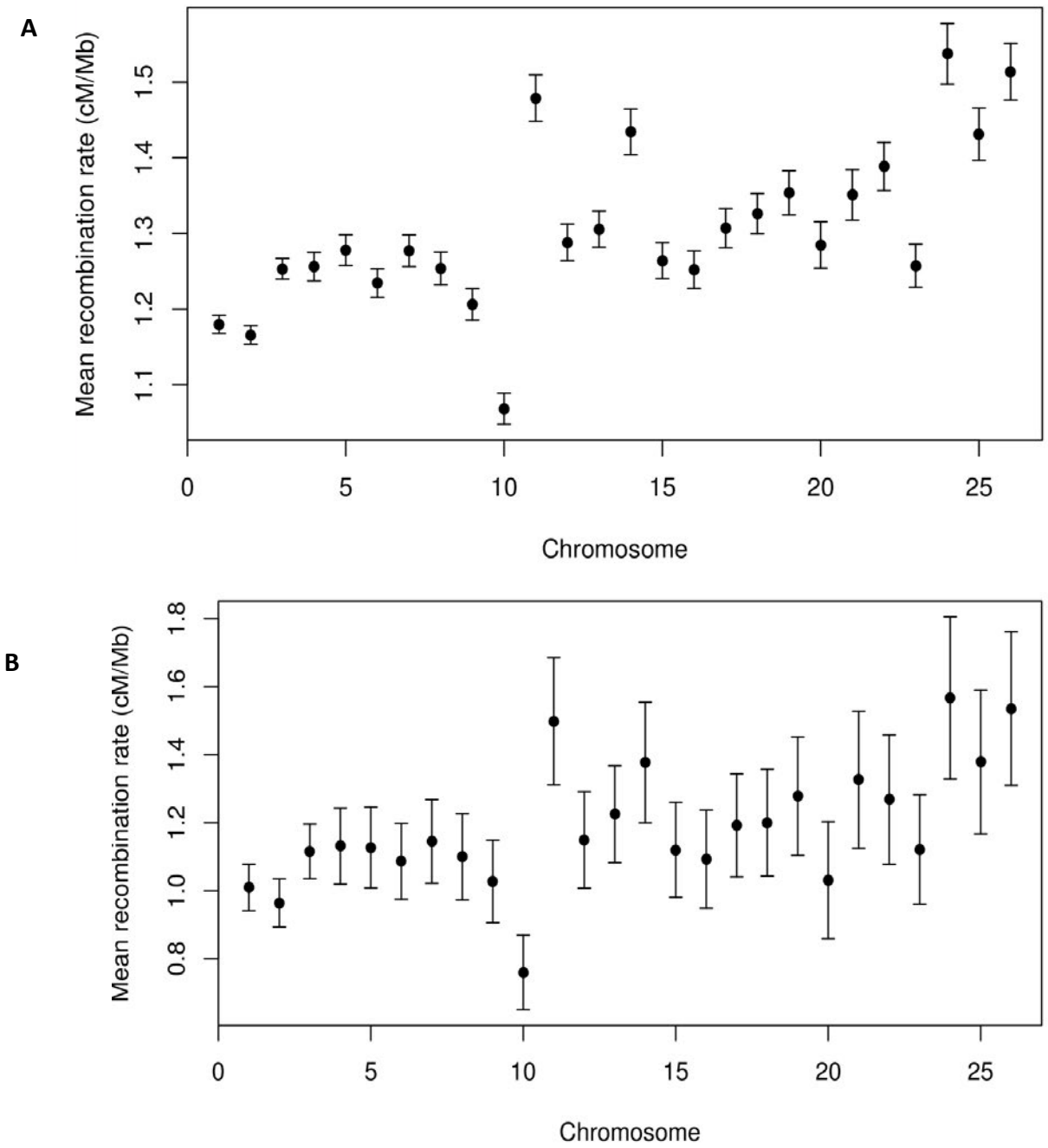

Figure 11 : Taux de recombinaison moyen de chaque autosome en Lacaune

A/ Taux de recombinaison estimés pour les chromosomes sur les intervalles de la 50K. **B/** Taux de recombinaison estimés pour les chromosomes sur des fenêtres de 1 Mb.

Sur la **Figure 11**, nous pouvons remarquer que, même lorsque la taille des chromosomes est prise en compte, il y en a quand même qui montrent un taux de recombinaison particulièrement

faible (comme les chromosomes 9, 10 ou 20), ou au contraire particulièrement élevé (c'est le cas pour les chromosomes 11 et 14). Les chromosomes présentant un faible taux de recombinaison possèdent de larges zones où la recombinaison est quasiment nulle ; entre 9 et 14 Mb pour le chromosome 9, entre 36 et 46 Mb pour le chromosome 10 et entre 27 et 31 Mb pour le chromosome 20. L'augmentation du taux de recombinaison pour les chromosomes avec un fort taux varie selon les chromosomes, ainsi la recombinaison est globalement augmentée sur le chromosome 14, alors que le chromosome 11 montre deux zones où la recombinaison est forte : entre 7 et 8 Mb et entre 53 et 54 Mb. De plus, dans les intervalles de la 50K, le taux de *GC* est positivement corrélé avec le taux de recombinaison (p-valeur $<10^{-16}$, corrélation r = 0,23). C'est également vrai pour les intervalles de 1 Mb (p-valeur $< 10^{-16}$ et corrélation r = 0,28). Ce taux de *GC* a pu être estimé en utilisant la séquence de référence du mouton. A partir de cette séquence, et dans les intervalles donnés, il a été possible de calculer les proportions des bases *G* et *C* par rapport à toutes les bases de l'intervalle, ce qui donne le taux de *GC*.

III. L'étude des cartes de recombinaison en données populationnelles

II. 4. *Méthode de détection des points chauds de recombinaison*

Afin de rechercher l'existence de potentiels points chauds en Lacaune, nous utilisons le deuxième jeu de données : les 51 Lacaune non-apparentés et non présents dans le jeu de données familial, génotypés pour la puce 600K. A l'aide de la méthode de Li et Stephens (2003), présentée dans le Chapitre 1 (« II. 1. a. b. »), et du logiciel PHASE développé par cette équipe, il est possible d'estimer les taux de recombinaison populationnels. Ces estimations ont été réalisées dans des fenêtres ***w*** de 2 Mb, pour des raisons calculatoires, avec ajout de 100 Kb de chaque côté des fenêtres chevauchant les fenêtres voisines, afin d'éviter les effets de bord lors des calculs effectués par PHASE. En effet, pour réaliser ses estimations, le logiciel utilise l'information des haplotypes étant à gauche et à droite de la fenêtre considérée, lorsque ce sont les fenêtres des extrémités des chromosomes, il est donc nécessaire d'ajouter de l'information afin de ne pas biaiser les estimations. Le logiciel PHASE a été lancé avec les options par défaut, nous avons juste

augmenté le nombre d'itérations afin d'obtenir de plus larges échantillons des taux de recombinaison dans la distribution *a posteriori*.

A partir des fichiers de sortie du logiciel, nous avons ainsi obtenu 1 000 échantillons issus de la distribution *a posteriori* pour :

- le taux de recombinaison populationnel de base : $\boldsymbol{\rho_w = 4*Nc_w}$, avec : ***N***, la taille efficace de la population, inconnue et qu'il n'est pas possible d'estimer, et $\boldsymbol{c_w}$, le taux de recombinaison dans les fenêtres ***w***, il est comparable à celui estimé en données familiales.

- une intensité de recombinaison spécifique $\boldsymbol{\lambda_j}$, qui a été définie au « Chapitre 1, II. 1. a. b », pour l'intervalle ***j*** de la puce 600K compris dans la fenêtre ***w***, sachant que la longueur de l'intervalle ***j*** est donnée par $\mathbf{l_j}$.

Il est donc possible d'estimer une distance génétique propre à la recombinaison populationnelle et à un intervalle ***j*** : $\boldsymbol{\delta_j}$, qui dépend de la recombinaison populationnelle de base $\boldsymbol{\rho_w}$, de la taille de l'intervalle $\boldsymbol{l_j}$, mais également de la recombinaison propre à l'intervalle, qui peut être différente de la recombinaison populationnelle de base et cette différence est exprimée par l'intensité $\boldsymbol{\lambda_j}$. Cette distance génétique s'exprime donc de la manière suivante :

$$\boldsymbol{\delta_j = \rho_w \lambda_j l_j}.$$

Il est intéressant de constater que la recombinaison populationnelle de base, $\boldsymbol{\rho_w}$, varie le long du génome (**voir Figure 12**).

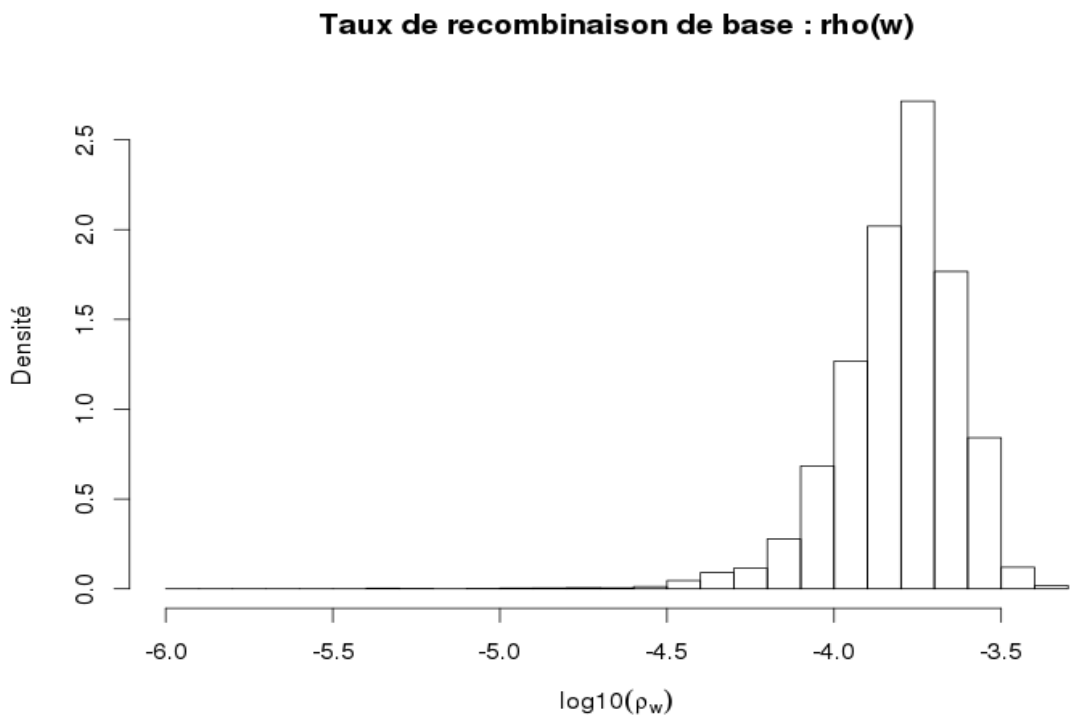

Figure 12 : Observation de la distribution de la recombinaison populationnelle de base le long du génome

Observation des log10 du taux de recombinaison populationnel de base $\boldsymbol{\rho_w}$ pour les fenêtres ***w*** de 2 Mb. Nous remarquons que le taux de recombinaison varie le long du génome.

Contrairement aux taux de recombinaison familiaux, les estimations faites en données populationnelles sont plus précises, car elles exploitent de nombreuses méioses accumulées sur plusieurs générations. Cependant, les taux de recombinaison populationnels dépendent de la taille efficace de la population, qui est inconnue ici, et qui peut varier le long du génome à cause de pressions évolutives, telles que la sélection. Les taux de recombinaison populationnels sont sensibles à ces phénomènes démographiques et ces pressions évolutives et peuvent donc être biaisés. De plus, ils ne permettent pas de connaître des taux individuels, uniquement ceux d'une population entière. Malgré tout, leur grande précision rend possible la détection de potentiels points chauds.

II. 5. *Les points chauds de recombinaison existent en Lacaune*

Afin de détecter de potentiels points chauds, nous utilisons les intensités de recombinaison spécifiques $\boldsymbol{\lambda_j}$ déterminées pour chaque intervalle ***j*** (voir précédemment). Les intervalles ***j*** montrant une intensité de recombinaison largement supérieure au taux de recombinaison populationnel de base dans cet intervalle sont considérés comme étant des points chauds. Plus particulièrement, nous avons représenté la distribution des ***log10(λj)*** tout génome (**voir Figure 13**), en faisant l'hypothèse que, sous H_0 (il n'y pas de points chauds), la distribution suivait une loi Normale et que $\lambda = 1$.

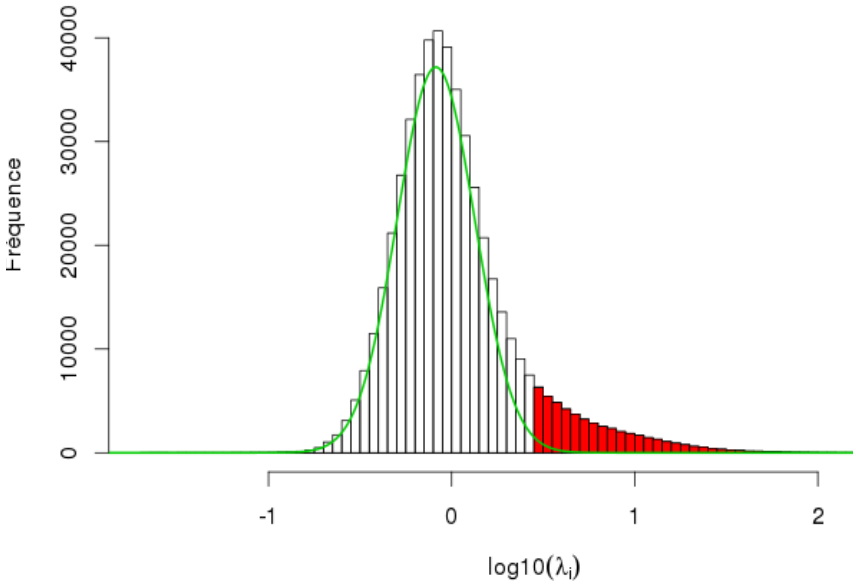

*Figure 13 : Distribution des intensités de recombinaison λ_j sur les intervalles **j** de la 600K*

La courbe en vert représente la distribution sous l'hypothèse nulle H_0, il n'y a pas de points chauds ($\log10(\lambda_j) = 0$). Les intervalles avec une intensité de recombinaison très importante, pour un *FDR* de 5 %, sont considérés comme des points chauds et sont illustrés en rouge.

Finalement, nous considérons qu'un intervalle est un point chaud lorsque le $FDR(\lambda_j) < 5$ %, sachant que le *FDR* (False Discovery Rate) est estimé avec la méthode de Storey et Tibshirani (2013), implémentée dans le package R « q-value ». Nous avons ainsi considéré environ 50 000 intervalles avec une intensité de recombinaison suffisante pour être des points chauds. Ils ont une distance médiane d'environ 40 Kb. En revanche, lorsque nous sommes plus stringent, avec un *FDR* de 0,1 %, nous détectons seulement 20 000 points chauds. Ces points chauds-là, plus précis,

permettront de rechercher des motifs d'*ADN* particuliers (*cf.* « VII. 2. »). En revanche, les 50 000 points chauds serviront surtout à étudier l'effet de la densité des points chauds sur le taux de recombinaison et pour cette étude, il n'est pas nécessaire d'être aussi sûr que pour la recherche de motifs.

Nous avons également cherché à savoir comment se distribuaient les crossing-overs sur le génome ; dans quelle proportion du génome a lieu la majorité de la recombinaison. Pour cela, nous calculons les distances génétiques cumulées, ainsi que les distances physiques cumulées pour les intervalles de la 600K et nous représentons les secondes en fonction des premières (**voir Figure 14**). Cela a permis de montrer que 80 % des crossing-overs ont lieu dans seulement 40 % du génome et que 60 % des crossing-overs ont lieu dans 20 % du génome.

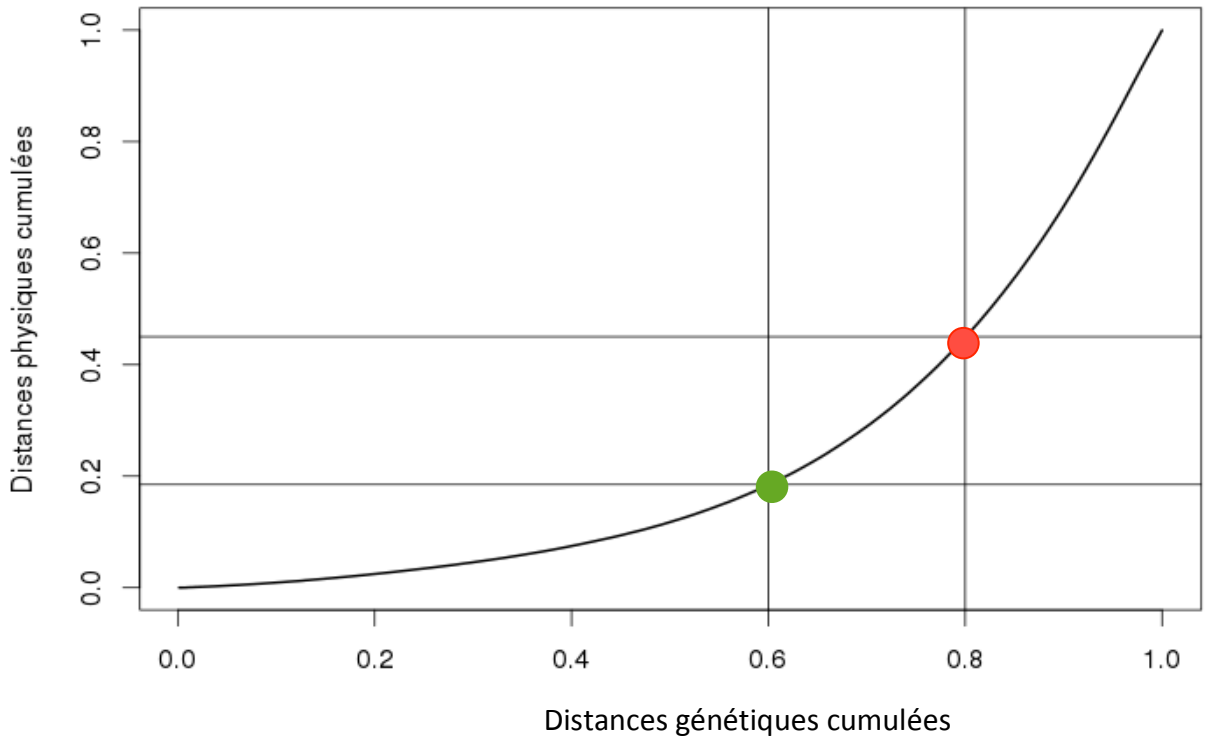

Figure 14 : Distribution de la recombinaison sur le génome

Représentation de la proportion de la taille physique du génome impactée par la recombinaison. En vert ; 60 % de la recombinaison a lieu dans 20 % du génome et en rouge ; 80 % de la recombinaison a lieu dans seulement 40 % du génome.

Une autre façon de représenter la proportion de recombinaison ayant lieu dans une certaine part du génome est l'utilisation du coefficient de Gini (Kaur et Rockman, 2014). Il prend des valeurs

comprises entre 0 (les crossing-overs sont distribués de manière homogène sur le génome) et 1 (tous les crossing-overs ont lieu dans une seule région du génome). Pour calculer ce coefficient, il faut, comme pour la figure précédente, trier les intervalles en fonction de leur taux de recombinaison, ce qui permet de tracer les distances physiques cumulées en fonction des distances génétiques cumulées (**voir Figure 15**). Si la recombinaison est constante sur le génome, la courbe sera proche de la droite y = x. Au contraire, plus le taux de recombinaison sera inégalement réparti sur le génome, plus la courbe va s'éloigner de la droite y = x et le coefficient de Gini va permettre d'estimer cette différence en mesurant l'aire sous la courbe : le coefficient étant égal à :

1 – (2*aire sous la courbe).

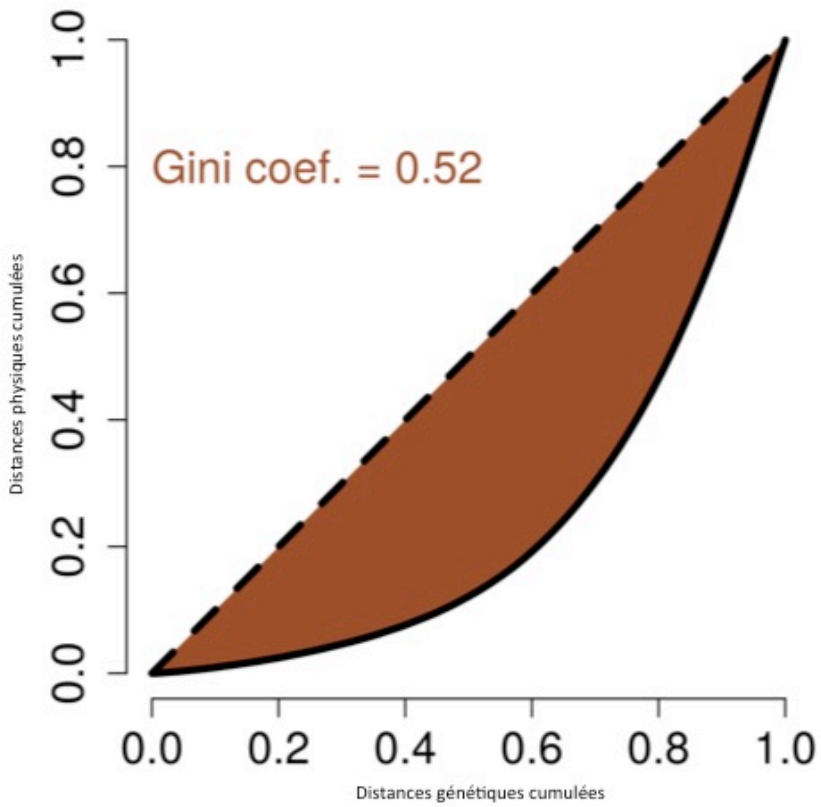

Figure 15 : Représentation de la distribution sur le génome à l'aide du coefficient de Gini

La figure représente la proportion de la taille physique du génome affectée par la recombinaison, pour une distance génétique qui augmente le long du génome. Le coefficient de Gini correspond à l'aire en marron sur la figure et vaut 0,52 en Lacaune.

En race Lacaune, nous avons ainsi obtenu un coefficient de Gini proche de 0,52. Ce coefficient est également intéressant car il permet de comparer les espèces entre elles, ce que nous ferons en discussion de cette partie.

III. La création de cartes génétiques de haute résolution grâce à la combinaison des deux jeux de données

III. 1. *Méthode d'obtention des cartes de haute résolution*

Afin de construire des cartes de recombinaison familiales sur les intervalles de la puce 600K, il est nécessaire d'échelonner le taux de recombinaison populationnel en le multipliant par 4 fois la taille efficace de la population. Or, à cause de pressions évolutives, la taille efficace de la population varie le long du génome et il faut donc pouvoir l'estimer localement. Cette estimation peut être réalisée en utilisant des échantillons des taux de recombinaison familiaux préalablement obtenus (***cj***).

Pour une fenêtre ***j*** de 1 Mb sur le génome, grâce à la même méthode que celle présentée précédemment (Chapitre 2, « II. 2. »), il est possible d'échantillonner ***k*** valeurs c_{jk} (issues de la fenêtre ***j*** et de l'échantillon ***k***) pour la distribution *a posteriori* du taux de recombinaison familial c_j, obtenu avec le jeu de données familiales. Puis, en réutilisant le logiciel PHASE, il est également possible d'extraire ***k*** échantillons ρ_{jk} issus de la distribution *a posteriori* des taux de recombinaison populationnels ρ_j, obtenus avec le jeu de données populationnel.

En considérant que :

$$\rho_j = \rho_w * \lambda_j \text{ c'est-à-dire, } \rho_j = 4N_j * c_j * \lambda_j \text{ avec :}$$

- N_j : taille efficace locale de la population dans la fenêtre ***j***.

λ étant une constante, il est ensuite possible d'obtenir :

$$\log(\rho_j) = \log(4N_j) + \log(c_j)$$

Ceci permet donc de construire le modèle suivant combinant les taux de recombinaison familiaux et populationnels, et prenant en compte un effet méthode ***i*** :

$$y_{ijk} = \mu + x_{ijk} * \alpha + \beta_j + v_{ij} + e_{ijk} \quad (4) \text{ avec :}$$

- $\mathbf{y_{ijk}}$ correspond au $\mathbf{log(c_{jk})}$ lorsque $\boldsymbol{i}$ = 1 (échantillons issus des taux de recombinaison familiaux) : $\mathbf{log(c_{jk}) = \mu + 0*\alpha + \beta_j + v_{1j} + e_{1jk}}$

- $\mathbf{y_{ijk}}$ correspond au $\mathbf{log(\rho_{jk})}$ lorsque $\boldsymbol{i}$ = 2 (échantillons issus des taux de recombinaison populationnels) : $\mathbf{log(\rho_{jk}) = \mu + 1*\alpha + \beta_j + v_{2j} + e_{2jk}}$

- $\boldsymbol{\mu}$: estimation du log du taux de recombinaison sur le génome entier.

- $\mathbf{x_{ijk}}$ = 1 si $\boldsymbol{i}$ vaut 2 et 0 sinon, si bien que $\boldsymbol{\alpha}$ estime le $\boldsymbol{log(4N)}$ où $\boldsymbol{N}$ est la taille efficace moyenne de la race Lacaune.

- $\boldsymbol{\mu + \beta_j}$: estimation des $\boldsymbol{log(c_j)}$ combinant les taux de recombinaison populationnels et familiaux.

- $\mathbf{\alpha + (v_{2j} - v1_j)}$: estimation des $\boldsymbol{log(4N_j)}$.

En effet, le calcul de la différence, en log, des taux de recombinaison familiaux et populationnels donne :

$\mathbf{log(c_j) - log(\rho_j) = -\alpha + (v_{1j} - v_{2j})}$, ce qui correspond bien à une variation locale, propre à la méthode, de la taille efficace de la population.

$\boldsymbol{\mu}$ et $\boldsymbol{\alpha}$ sont considérés comme des effets fixes, tandis que $\boldsymbol{\beta_j}$ et $\boldsymbol{v_{ij}}$ sont considérés comme des effets aléatoires indépendants mais qui viennent d'une même distribution de loi Normale. L'utilisation de cette approche permet de combiner dans un seul modèle les estimations issues des jeux de données populationnels et familiaux, tout en considérant leurs incertitudes respectives puisque nous exploitons les distributions *a posteriori*.

Dans la thèse, nous appliquons l'équation **(4)** avec un effet fixe supplémentaire, correspondant au chromosome, en utilisant le package lme4 de R (Bates *et al.*, 2015). Nous considérons des fenêtres de 1 Mb couvrant l'intégralité du génome, grâce à 20 échantillons issus des distributions *a posteriori* de $\boldsymbol{c_j}$ et $\boldsymbol{\rho_j}$. Nous ne prenons pas en compte les fenêtres situées à moins de 4 Mb de chaque extrémité des chromosomes, en raison de biais dans l'estimation des $\boldsymbol{c_j}$ dans ces régions. Après estimation de ce modèle, nous corrigeons les taux de recombinaison populationnels de chaque intervalle de la 600K en les divisant par la taille efficace locale de la population $\boldsymbol{N_j}$. En revanche, comme nous voulons quand même une carte sur tout le génome, pour les intervalles situés dans les 4 Mb de chaque extrémité des chromosomes, les taux de recombinaison populationnels sont corrigés par la taille efficace moyenne de la population $\boldsymbol{N}$. Cela

permet d'estimer des taux de recombinaison, exprimés en cM/Mb, pour chaque intervalle de la puce 600K et donc d'obtenir une carte de recombinaison haute résolution.

III. 2. *Impact des points chauds sur le taux de recombinaison*

La construction d'une carte de recombinaison haute résolution à l'aide de deux jeux de données indépendants, cependant issus de la même race, la Lacaune, permet d'observer l'effet des points chauds sur le taux de recombinaison.

Pour cela, nous comptons le nombre de points chauds significatifs pour un *FDR* de 5 % dans chaque intervalle de la puce 50K (identifiés dans la partie II.5). Après avoir corrigé pour le chromosome, le taux de *GC*, et après avoir supprimé les régions de 4 Mb à chaque extrémité des chromosomes, nous appliquons une régression linéaire sur le taux de recombinaison calculé dans les intervalles, afin d'estimer l'effet du nombre de points chauds et de la densité des points chauds, qui est définie comme le nombre de points chauds dans l'intervalle, divisé par la taille de l'intervalle.

A la suite de cette régression linéaire, il apparaît que le nombre de points chauds et la densité de points chauds sont très fortement associés au taux de recombinaison familiaux : corrélation r = 0,15 et p-valeur < 10^{-16} pour la densité en points chauds, et corrélation r = 0,19 et p-valeur < 10^{-16} pour le nombre de points chauds. Ces corrélations ont été obtenues après avoir corrigé pour l'effet des chromosomes et des taux de *GC* sur la recombinaison : corrélation r = 0,14 et p-valeur < 10^{-16} et corrélation r = 0,18 et p-valeur < 10^{-16} respectivement. La **Figure 16** illustre ce résultat dans deux fenêtres de 1 Mb, provenant du chromosome 24. La première, entre 5 et 6 Mb, présente un fort taux de recombinaison (7,08 cM/Mb), la deuxième, entre 18 et 19 Mb, un taux plus faible (0,46 cM/Mb). Lorsque les deux fenêtres sont comparées, nous remarquons que celle qui recombine beaucoup possède 36 points chauds, alors qu'il n'y en a aucun dans la fenêtre qui recombine peu. Etant donné que le taux de recombinaison populationnel de base est très similaire entre les deux fenêtres (0,7/Kb pour la fortement recombinante, 0,2/Kb pour la deuxième, le taux de recombinaison populationnel n'a pas d'unité), la différence locale du taux de recombinaison entre les deux fenêtres est principalement due à leur nombre de points chauds respectifs.

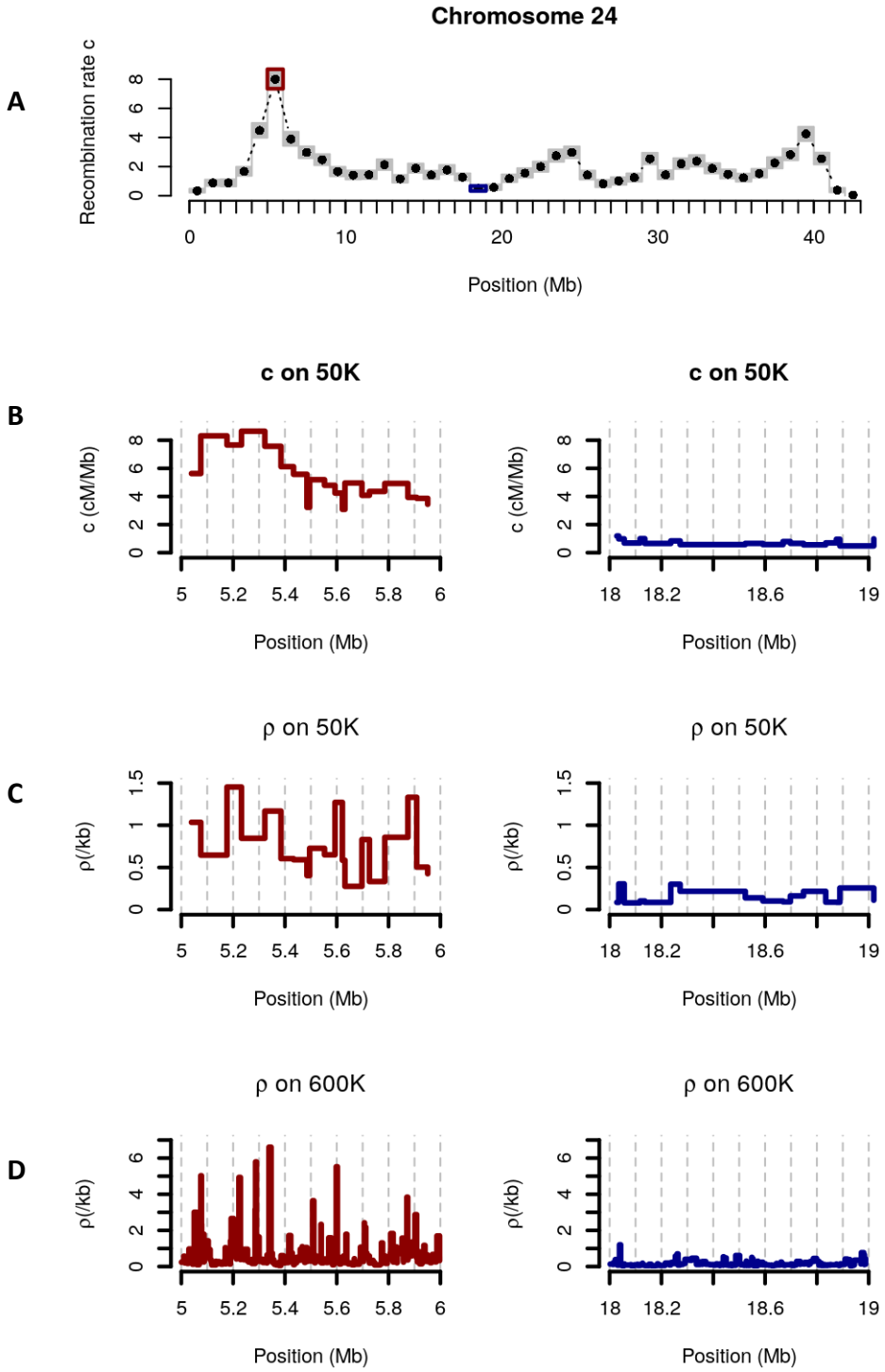

Figure 16 : Effet des points chauds sur le taux de recombinaison familial

Zoom sur deux fenêtres du chromosome 24 ; l'une qui recombine beaucoup (à gauche, en rouge), l'autre qui recombine beaucoup moins (à droite, en bleu). **A/** Estimation des taux de recombinaison familiaux ***c*** dans les intervalles de la puce 50K. **B/** Estimation des taux de recombinaison populationnels ***ρ***

dans les intervalles de la puce 50K. **C/** Estimation des taux de recombinaison populationnels ***ρ*** dans les intervalles de la puce 600K.

III. 3. *Détection de signatures de sélection*

La carte de recombinaison haute résolution peut aussi être utile pour estimer l'impact des pressions évolutives sur la race Lacaune. Le modèle linéaire présenté a permis d'estimer la taille efficace moyenne de la race Lacaune, la corrélation entre les taux de recombinaison populationnels et familiaux et d'identifier des régions du génome où les deux types de taux de recombinaison étaient significativement différents.

La taille efficace de la race Lacaune a donc été estimée autour de 7 000 individus et une corrélation de 73 % a été déterminée entre le taux de recombinaison populationnel et le taux de recombinaison familial (**voir Figure 17**).

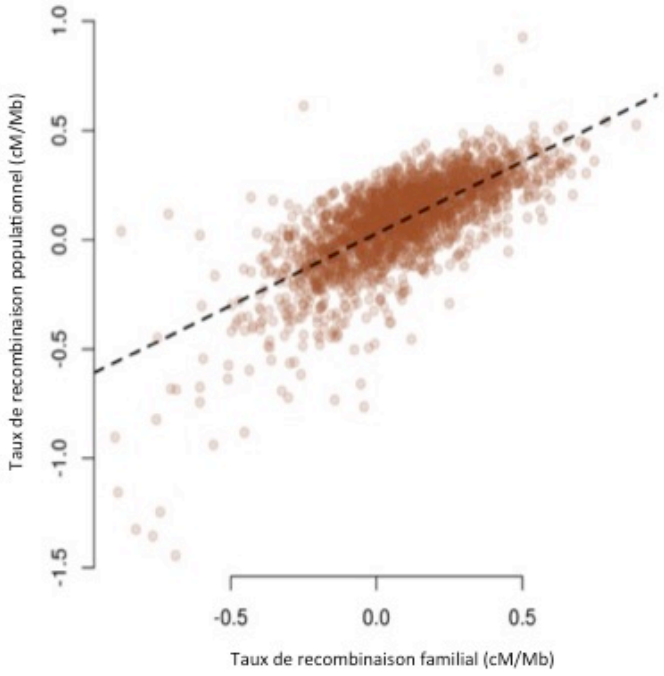

Figure 17 : Taux de recombinaison populationnel et familial dans des fenêtres de 1 Mb

La ligne en pointillé représente la régression du taux de recombinaison populationnel sur le taux de recombinaison familial. Les valeurs sont montrées sur une échelle logarithmique.

Comme annoncé précédemment, le taux de recombinaison populationnel peut être sensible aux pressions évolutives. Ainsi lorsque nous regardons la **Figure 12** (partie II.4), nous remarquons que, tout à gauche, les taux de recombinaison sont très faibles. Or, cela pourrait être dû à des signatures de sélection qui diminuent plus ou moins fortement le taux de recombinaison populationnel. Par exemple, sur la **Figure 18**, le point rouge illustre une potentielle trace de sélection identifiée préalablement sur le chromosome 6 par Rochus *et al.* (2017) : le taux de recombinaison familial $\boldsymbol{c_w}$ est normal, en revanche, le taux de recombinaison populationnel $\boldsymbol{\rho_w}$ est très faible.

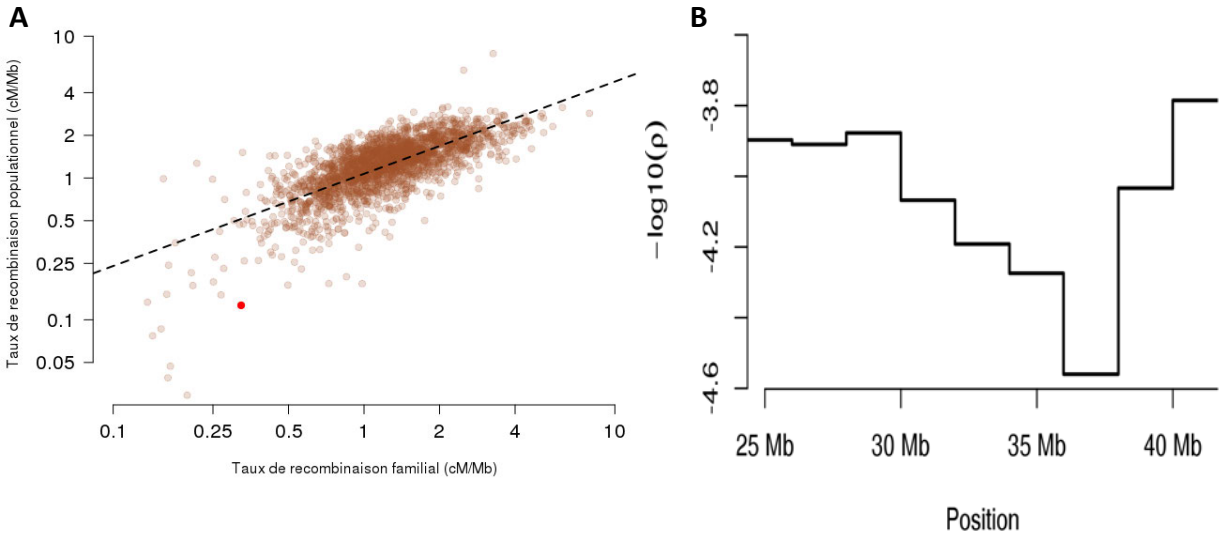

Figure 18 : Impact des signatures de sélection sur le taux de recombinaison populationnel

A/ Taux de recombinaison populationnel en fonction du taux de recombinaison familial. Le point rouge représente une potentielle trace de sélection sur le chromosome 6 : le taux de recombinaison familial est proche de 0,3, alors que le taux de recombinaison populationnel reste très faible. **B/** Zoom sur le chromosome 6 et sur la trace de sélection découverte (Rochus *et al.*, 2017) : la baisse du taux de recombinaison populationnel $\boldsymbol{\rho}$ est bien visible.

L'étude de la carte de recombinaison haute densité a permis de mettre en évidence 7 régions où le taux de recombinaison populationnel est beaucoup plus faible que le taux de recombinaison familial et 3 régions où c'est l'inverse (**voir Figure 19**). Sept des 10 régions ont des taux de recombinaison familiaux extrêmes par rapport au reste du génome (**voir Tableau 3**). Pour quantifier à quel point ces régions $\boldsymbol{w}$ sont extrêmes, la proportion du génome, $\boldsymbol{q_w}$, où le taux de

recombinaison familial est faible a été calculée ; pour des régions où le taux de recombinaison familial est supérieur au taux de recombinaison populationnel, plus q_w est proche de 0, plus les taux de recombinaison familiaux sont différents du reste du génome, à l'inverse pour les régions où le taux de recombinaison populationnel est supérieur au taux de recombinaison familial, plus q_w est proche de 1, plus les régions sont extrêmes. Cela permet de voir si les faibles taux de recombinaison sont effectivement dus à des signatures de sélection ou à des biais causés par les méthodes d'estimation.

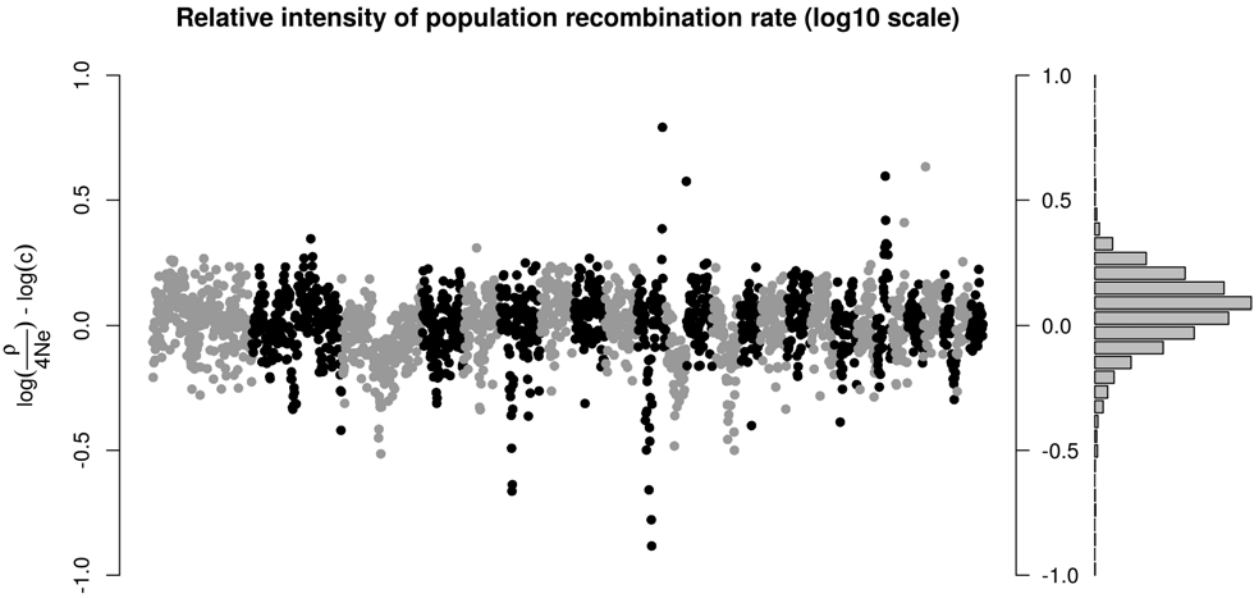

Figure 19 : Intensité relative du taux de recombinaison populationnel par rapport au taux de recombinaison familial dans des fenêtres de 1 Mb

Utilisation d'un modèle linéaire mixte permettant d'estimer la taille efficace de la race Lacaune et l'observation de régions où les taux de recombinaison familiaux et populationnels diffèrent. Un changement de couleur correspond à un chromosome. Observation de 7 régions où ρ est très inférieur à c et 3 régions où ρ est très supérieur à c.

Pour 6 régions sur les 7 ayant un taux de recombinaison familial extrême, le taux de recombinaison populationnel est encore plus extrême que le taux de recombinaison familial (**voir Tableau 3**) : 4 régions ont ainsi un faible taux de recombinaison familial, et un taux de recombinaison populationnel encore plus faible (les deux régions du chromosome 3 qui ont des q_w de 0,06 et de 0,04 respectivement et deux régions du chromosome 10, entre 36 et 37 Mb et entre

42 et 44 Mb, qui ont des $\boldsymbol{q_w}$ de 0,01 et de moins de 0,01 respectivement). Deux autres régions ont à l'inverse un taux de recombinaison familial très élevé et un taux de recombinaison populationnel encore plus fort (sur les chromosomes 12 et 23 où les $\boldsymbol{q_w}$ valent 0,92 et 0,97 respectivement). Pour ces 6 régions, la différence entre le taux de recombinaison familial et le taux de recombinaison populationnel peut être expliquée par le modèle utilisé pour l'estimation du taux de recombinaison familial, qui a tendance à ramener les estimations vers la moyenne. Or, le taux de recombinaison populationnel n'est pas ramené vers la moyenne de la même façon et il est donc possible que les taux de recombinaison estimés pour les 6 régions ne concourent pas et que leur taux de recombinaison familial soit sous-(ou sur-)exprimé.

Sur les 4 régions restantes, trois ont un taux de recombinaison populationnel faible, mais leur taux de recombinaison familial n'est pas spécialement extrême ; la différence ne semble donc pas, ici, être due à un problème de méthode. De plus, ces 3 régions correspondent à des signatures de sélection qui on été préalablement identifiées chez le mouton : une région sur le chromosome 6 qui couvrent 2 intervalles entre 36 et 38 Mb et qui contient le gène *ABCG2*, associé à la production laitière (Cohen-Zinder *et al.*, 2005), et le gène *LCORL*, associé à la stature (Takasuga, 2015). Elle correspond à la région présentée précédemment (**voir Figure 18**), qui est sous sélection en race Lacaune (Fariello *et al.*, 2014, Rochus *et al.*, 2017). Une autre région couvre un intervalle sur le chromosome 10, entre 29 et 30 Mb, et contient le gène *RXFP2*, associé aux phénotypes cornu et sans cornes (Johnston *et al.*, 2013) et qui est sous sélection dans de nombreuses races ovines (Fariello *et al.*, 2014). La dernière région se situe sur le chromosome 13, entre 63 et 64 Mb, et contient le gène *ASIP* responsable de la couleur de la robe de certains moutons (Norris et Whan, 2008) et est, là-encore, sous sélection. Pour ces trois régions, la baisse du taux de recombinaison populationnel peut être due à une réduction locale de la taille efficace de la population, causée par la sélection.

Tableau 3 : Régions du génome où les taux de recombinaison populationnel et familial sont significativement différents

Chromosome	Taille des intervalles (Mb)	q_w	P-valeur	ρ/c
3	103-104	0,06	$1,6*10^{-5}$	0,28
3	109-110	0,04	$1,8*10^{-5}$	0,28
6	**36-38**	**0,14**	**$1,2*10^{-7}$**	**0,22**
10	**29-30**	**0,77**	**$8,8*10^{-5}$**	**0,31**
10	36-37	0,01	$2,1*10^{-5}$	0,29
10	42-44	< 0,01	$1,2*10^{-14}$	0,11
13	**63-64**	**0,33**	**$7,4*10^{-6}$**	**0,31**
12	4-5	0,92	$7,4*10^{-6}$	3,7
20	**28-29**	**0,01**	**$1,7*10^{-5}$**	**3,6**
23	10-11	0,97	$5,1*10^{-6}$	3,8

$\boldsymbol{\rho}$: Taux de recombinaison populationnel de base, $\boldsymbol{c}$: taux de recombinaison familial. $\boldsymbol{\rho/c}$: rapport du taux de recombinaison populationnel sur le taux de recombinaison familial. $\boldsymbol{q_w}$: proportion du génome avec un faible taux de recombinaison familial. Les régions avec une p-valeur < 10^{-4} sont considérées comme présentant des différences significatives pour un *FDR* de 2 %. Les régions en gras correspondent à de potentielles signatures de sélection.

Finalement, sur les 3 régions présentant un taux de recombinaison populationnel supérieur au taux de recombinaison familial, une est située sur le chromosome 20, entre 28 et 29 Mb. Etant donné que son taux de recombinaison familial n'est pas élevé, l'effet de la méthode ne semble pas s'appliquer ici. Cette région contient un groupe de gènes codant pour des récepteurs olfactifs. Elle pourrait ainsi être expliquée comme ayant subi des pressions de sélection afin d'augmenter la diversité génétique, ce qui a déjà pu être vu dans d'autres espèces, telles que le porc, l'humain ou les rongeurs (Groenen *et al.*, 2012, Ignatieva *et al.*, 2014, Stathopoulos *et al.*, 2014).

En plus de la zone entre 29 et 30 Mb sur le chromosome 10, il y a également une zone assez large, entre 30 et 45 Mb, où la recombinaison populationnelle est quasiment nulle. Cette large

zone pourrait expliquer en partie le très faible taux de recombinaison observé sur le chromosome 10 (**voir Figure 11**). Cette zone est très riche en *AT* et semble de plus conservée entre les espèces, car chez l'Homme aussi elle recombine très peu.

IV. Comparaison des cartes génétiques entre Lacaune et Soay

Comme je l'ai indiqué précédemment, la seule étude du déterminisme génétique de la variation inter-individuelle de la recombinaison en mouton à ce jour a été faite en race Soay (Johnston *et al.*, 2016). Etant donné que leurs données et résultats étaient disponibles, il était intéressant de pouvoir les comparer avec les nôtres (**voir Tableau 4**). L'étude en Soay a été réalisée sur les mâles et les femelles, cependant, comme nous n'avons pu étudier la recombinaison que chez les Lacaune mâles, nous n'avons utilisé que les données disponibles en Soay mâle pour la comparaison des cartes génétiques.

La race Soay est une population férale issue de la première vague de domestication des ovins et fait partie du groupe des races du Nord (**voir Figure 20**). Elle est majoritairement présente sur une île située au Nord-Ouest de l'Ecosse et est plutôt soumise à de la sélection naturelle. La Lacaune, en revanche, fait partie du groupe des races du Sud et est issue de la deuxième vague de domestication. Les deux races sont donc très distantes génétiquement parlant et leur calcul de Fst, réalisé en utilisant les données de « SheepHapmap » (Kijas *et al.*, 2012) donne une valeur de 0,4. Le Fst, ou indice de différenciation, permet de mesurer la différenciation entre des populations à partir du polymorphisme génétique. En général, il est calculé en utilisant des *SNPs*.

Tableau 4 : Comparaison des données disponibles en Soay mâles et en Lacaune mâles

Race	Soay mâles	Lacaune mâles
Méioses	1 196	5 940
Nombre de pères	227	345
Nombre total de crossing-overs	40 807	213 615
Nombre moyen de crossing-overs	34,12	35,95
Nombre de marqueurs (issus de la 50K)	39 104	46 813

Il est intéressant de remarquer qu'il y a beaucoup moins de crossing-overs disponibles en Soay mâle qu'en Lacaune, ce qui peut s'expliquer par un manque d'informativité, dû à un nombre de descendants beaucoup plus faible. De plus, nous remarquons que le nombre de marqueurs utilisés est également plus faible, or la race Soay est beaucoup plus consanguine (**voir Figure 20**), ce qui conduit à moins de polymorphismes au niveau des *SNPs*.

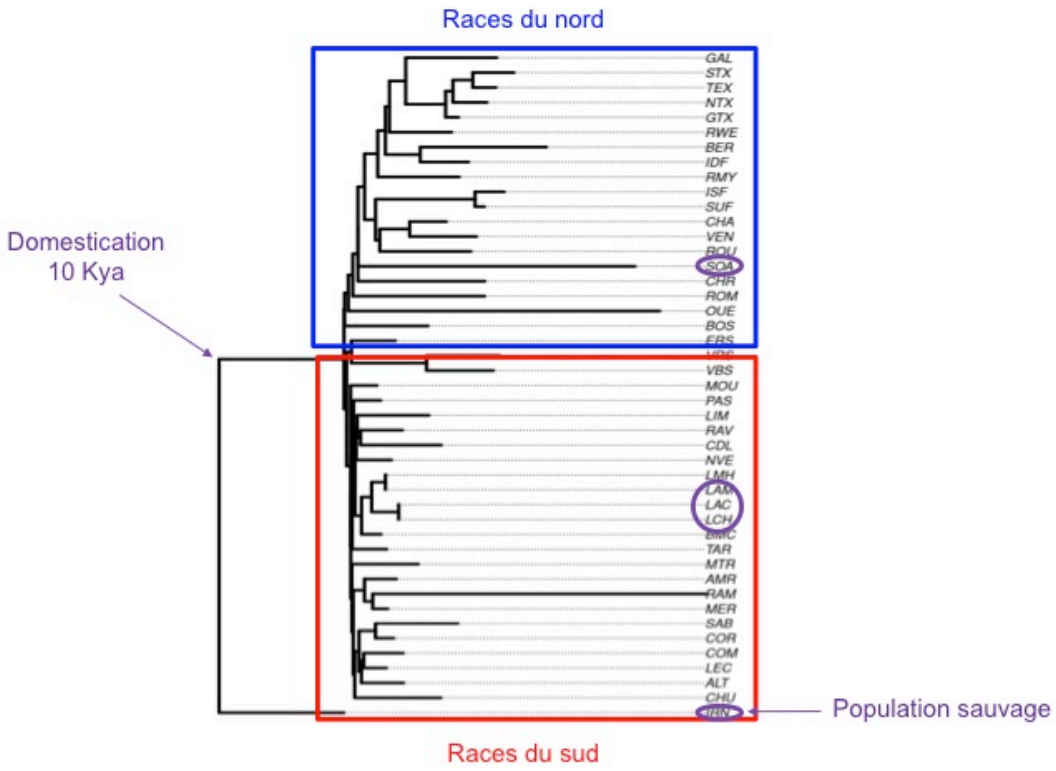

Figure 20 : Arbre phylogénétique des différentes races de mouton (d'après Rochus et al., 2017)

L'encadré bleu montre le groupe des races du nord, tandis que l'encadré rouge montre le groupe des races du sud. Le premier cercle violet indique la race Soay, le deuxième les deux rameaux Lacaune ; Lacaune viande et Lacaune lait, le troisième la population sauvage, les mouflons d'Iran. La longueur des branches indique le niveau de consanguinité des races ; plus elles sont longues, plus la race est consanguine.

Afin de pouvoir comparer les cartes de recombinaison obtenues en Lacaune avec celles des Soay, nous avons utilisé les données mises à disposition par l'équipe de Johnston *et al.* (2016). Cependant, comme ils n'ont pas utilisé la même méthode pour estimer les taux de recombinaison, nous les avons réestimés en utilisant notre méthode (présentée précédemment), afin que la comparaison ne soit pas impactée par des différences dues aux méthodes utilisées. Etant donné que notre jeu de données en Lacaune ne comporte que des mâles, nous n'avons construit des cartes génétiques que pour les Soay mâles également. Le jeu de données utilisé en Soay comprenait donc 3 445 animaux, dont 299 pères, similaires à nos 345 pères Lacaune. Nous avons ensuite appliqué LINKPHASE pour détecter les crossing-overs et nous avons obtenu 88 683

crossing-overs dans 2 609 méioses mâles. A la suite de cela, le taux de recombinaison familial a été estimé dans les mêmes intervalles que pour les Lacaune et de la même manière que présenté précédemment.

Cette comparaison a permis de montrer que les Lacaune et les Soay présentent des taux de recombinaison très similaires sur les intervalles de la 50K (**voir Figure 21**), avec une corrélation de 82 % (p-valeur < à 10^{-16}).

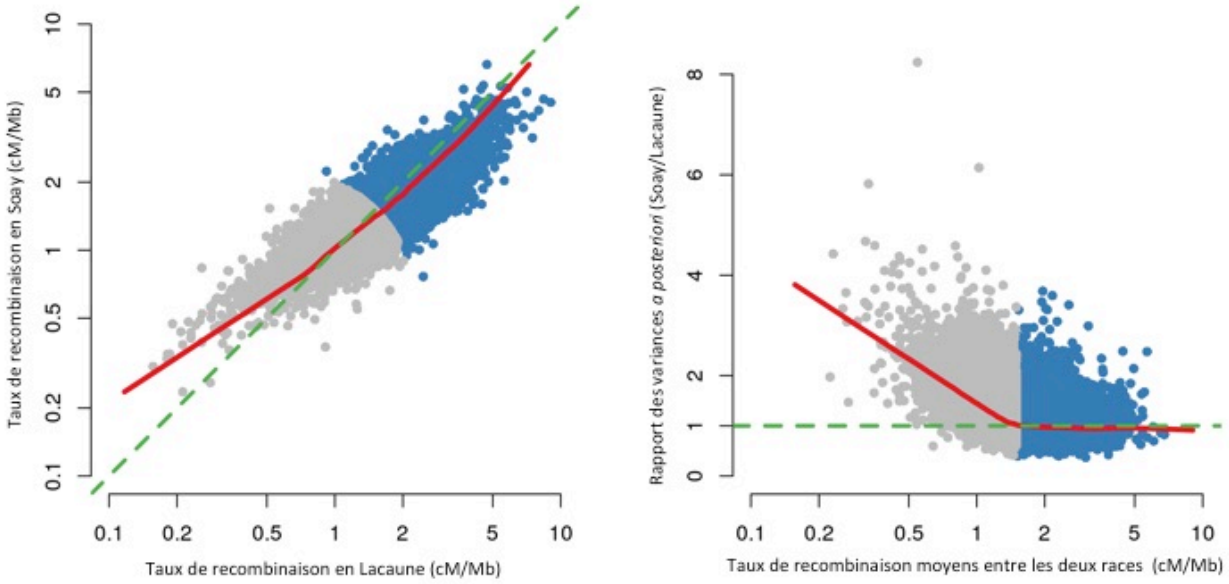

Figure 21 : Comparaison des taux de recombinaison entre les Lacaune et les Soay

A gauche : représentation en échelle log des taux de recombinaison moyens *a posteriori*. La droite verte correspond à la droite y = x et la droite rouge est une courbe lissée. A droite : représentation en échelle log du rapport des variances *a posteriori* (Soay/Lacaune) en fonction des taux de recombinaison moyens *a posteriori* dans les deux races. La droite verte correspond aux variances égales et la droite rouge est une courbe lissée. Les points gris sur les deux figures représentent des intervalles où le taux de recombinaison moyen est inférieur à 1,5 cM/Mb. Les moyennes et les variances *a posteriori* sont issues de notre modèle d'estimation du taux de recombinaison familial présenté au « Chapitre 2. II. 2. ».

Il est possible d'observer que les Soay présentent des taux de recombinaison plus forts pour des intervalles où le taux de recombinaison est faible (moins de 1,5 cM/Mb). Ceci peut s'expliquer par les effets du *prior* qui a tendance à ramener les estimations vers la moyenne. Cet effet serait plus prononcé chez les Soay car leur jeu de données est plus petit. En effet, le graphe de droite sur la **Figure 21** montre bien que les variances *a posteriori* du taux de recombinaison sont beaucoup plus

importantes en Soay pour des intervalles recombinant faiblement, alors qu'elles sont similaires sur le reste du génome.

Ces résultats indiquent bien que les deux races ont une même amplitude et une même distribution de la recombinaison sur le génome. Il nous a donc été possible de combiner les crossing-overs des deux races afin de créer une nouvelle carte de recombinaison, améliorée car avec plus d'informations, pour les intervalles de la 50K et des intervalles de 1 Mb. Nous avons ainsi obtenu 302 298 crossing-overs dans 8 549 méioses. Utiliser les deux populations permet de diminuer très clairement les variances des taux de recombinaison *a posteriori* et donc d'augmenter la précision des cartes.

V. Discussion

Au cours de cette thèse, nous avons pu créer des cartes fines de recombinaison pour deux résolutions différentes : une carte moyenne densité, grâce à la puce 50K et un large pédigrée et une carte haute densité, grâce à la puce 600K et 51 Lacaune non apparentés. Ce dernier jeu de données a également permis de détecter des points chauds. De plus, en combinant ces deux jeux de données, nous avons pu estimer la taille efficace de la race Lacaune et découvrir des signatures de sélection, certaines déjà connues, d'autres nouvelles.

Notre étude sur le taux de recombinaison familial à partir d'un large pédigrée en Lacaune a permis de montrer que la Lacaune présente des profils similaires à ceux des autres Mammifères (Shifman *et al.*, 2006, Chowdhury *et al.*, 2009, Tortereau *et al.*, 2012) : augmentation du taux de recombinaison aux extrémités des chromosomes, qu'ils soient acrocentriques ou métacentriques, diminution au niveau du centromère ou encore le taux de recombinaison qui augmente lorsque la taille des chromosomes diminue ce qui est cohérent avec le principe d'un crossing-over obligatoire par méiose.

V. 1. *La détection des points chauds de recombinaison*

L'étude de la carte de recombinaison obtenue avec les données populationnelles révèle des patrons de recombinaison à l'échelle de la kilobase : des zones du génome de très petite taille qui recombinent énormément et qui sont séparées par des zones plus larges et qui recombinent

beaucoup moins. Dans ces intervalles hautement recombinants, nous avons ainsi pu mettre en évidence des points chauds de recombinaison et nous avons observé que la majorité de la recombinaison avait lieu dans des petites portions du génome : nous avons ainsi estimé que 80 % de la recombinaison avait lieu dans 40 % du génome. L'utilisation d'un coefficient de Gini (Kaur et Rockman, 2014) permet de comparer cette distribution entre différentes espèces. En Lacaune, le coefficient obtenu vaut 0,52, ce qui est similaire à celui obtenu chez la drosophile, mais plus faible que le coefficient calculé chez l'Homme ou la souris (Kaur et Rockman, 2014). Cette proportion est cependant largement sous-estimée en raison de notre faible résolution (quelques Kb sur la puce 600K), alors que les points chauds ont plutôt une largeur de quelques centaines de paires de bases.

Malgré tout, nous avons identifié 50 000 points chauds significatifs pour un *FDR* de 5 % et 20 000 pour un *FDR* de 0,1 %. L'obtention de 50 000 points chauds est quasiment le double de ce qui a été estimé chez l'Homme (International HapMap Consortium *et al.*, 2007). Cette différence peut avoir différentes causes. Tout d'abord, il est possible que nos régions identifiées comme des points chauds de recombinaison soient dues à des erreurs d'assemblage du génome et nous avons effectivement un effet significatif, quoique modéré, (Odds Ratio = 1,4) de la présence de trous dans l'assemblage conduisant à détecter des points chauds là où il n'y en a pas. Cet Odds Ratio permet de calculer le rapport du nombre de points chauds sachant qu'il y a des trous dans l'assemblage sur le nombre de « non » points chauds sachant qu'il y a des trous dans l'assemblage et de le diviser par le rapport du nombre de points chauds sachant qu'il n'y a pas de trous dans l'assemblage sur le nombre de « non » points chauds sachant qu'il n'y a pas de trous dans l'assemblage. Cela permet de conclure qu'il y a 40% de chance d'avoir un point chaud dans un trou de l'assemblage. Deuxièmement, le test de significativité utilisé pour détecter nos 50 000 points chauds pourrait être trop permitif, en effet, lorsque nous utilisons un seuil plus stringent, nous obtenons environ 20 000 points chauds, ce qui est plus similaire à ce qui est trouvé chez l'Homme. Troisièmement, il a déjà été démontré que la sélection pouvait impacter la détection des points chauds, bien que normalement la méthode que nous avons utilisée devrait corriger l'impact éventuel de la sélection (Chan *et al.,* 2012). Pour finir, il est possible que les moutons ancestraux présentent plus de points chauds que les humains. Dans tous les cas, la forte corrélation entre le taux de recombinaison familial et la densité en points chauds montre que notre carte de

recombinaison populationnelle est relativement précise. Il sera possible d'obtenir une meilleure résolution pour détecter les points chauds en utilisant des données de séquençage.

V. 2. *La combinaison des deux jeux de données populationnel et familial*

Grâce à la création d'un modèle statistique, nous avons pu combiner les taux de recombinaison familiaux et populationnels. Utiliser une approche similaire à celle d'O'Reilly *et al.* (2008) a permis d'étudier les impacts des phénomènes démographiques et des pressions évolutives. Plus particulièrement, nous avons potentiellement découvert un signal de sélection diversifiante au niveau de récepteurs olfactifs. Cette région étant riche en gènes codant pour des récepteurs olfactifs, il est possible qu'une sélection diversifiante ait eu lieu historiquement dans cette région due à une augmentation de la taille efficace de la population et de la diversité génétique, afin de favoriser l'adaptation des moutons à leur milieu et leur permettre de mieux surveiller leur environnement, ces animaux étant considérés comme des proies dans la nature. Cela a notamment été observé chez le rat-taupe africain (Stathopoulos *et al.*, 2014). En effet, cet animal vivant sous terre, l'efficacité de son odorat est vue comme primordiale pour sa bonne évolution, ce qui se traduit par une sélection positive sur la grande diversité de gènes codant pour les récepteurs olfactifs.

La combinaison des deux types de recombinaison a révélé une corrélation d'environ 70 %, ce qui est plutôt bon, mais moindre que celle obtenue chez l'Homme : proche de 97 % sur des intervalles de 5 Mb (Myers *et al.*, 2006). En revanche, elle est proche de la corrélation calculée chez le ver, la souris ou la drosophile ; respectivement 69 %, 47 % et 50 % (Rockman et Kruglyak, 2009, Brunschwig *et al.*, 2012, Chan *et al.*, 2012). Là-encore, des estimations plus précises de la recombinaison familiale et de la recombinaison populationnelle pourraient affiner notre corrélation.

Ces différences entre espèces peuvent être expliquées par la méthode utilisée pour estimer le taux de recombinaison populationnel qui est basée sur le fait que la taille efficace de la population est constante, à la fois dans le passé et sur le génome. Pour prendre en compte la variation de la taille efficace, nous avons estimé le taux de recombinaison populationnel localement dans des intervalles de 2 Mb, mais il est malgré tout possible que la variation de la taille de la population

dans le passé affecte notre estimation du taux de recombinaison populationnel. En effet, la méthode a déjà été montrée comme étant affectée par la démographie, bien que ce soit moins le cas pour l'identification des points chauds (Li et Stephens, 2003). De plus, comme déjà mentionné précédemment, la sélection peut avoir un effet sur l'estimation du taux de recombinaison populationnel dans d'autres méthodes (Chan *et al.*, 2012), bien que cela n'ait pas été observé dans la méthode de Li et Stephens (2003).

Nos cartes de recombinaison familiales sont également uniquement basées sur les méioses mâles, or le taux de recombinaison populationnel est estimé à la fois avec des méioses mâles et des méioses femelles. Le fait que les taux de recombinaison diffèrent entre les mâles et les femelles (Johnston *et al.*, 2016) pourrait également expliquer cette plus faible corrélation.

De plus, la pression de sélection intense causée par la domestication, puis par la sélection artificielle, a pu modifier les patrons de déséquilibre de liaison sur le génome du mouton, ce qui a pour conséquence de diminuer la corrélation entre les deux recombinaisons. En effet, l'estimation de la recombinaison populationnelle résume la recombinaison qui a eu lieu dans le passé et il est donc possible que les points chauds de recombinaison, qui étaient présents dans la population ancestrale, ne soient aujourd'hui plus en activité chez les individus Lacaune. Ceci peut être le cas lorsque la domestication conduit à une réduction de la diversité des points chauds définis par un seul gène, comme *PRDM9*. En effet, une réduction du nombre de motifs d'*ADN* présents sous les points chauds pourrait conduire à une modification de la recombinaison le long du génome. Cela a été montré chez l'Homme où les patrons de recombinaison diffèrent entre des populations du fait de leur différence de diversité au niveau de *PRDM9* (Baudat *et al.*, 2010, Berge *et al.,* 2010, 2011). A terme, de tels phénomènes pourraient conduire à la dégradation de la corrélation entre la recombinaison actuelle (mesurée par le taux de recombinaison familial) et la recombinaison passée (mesurée par le taux de recombinaison populationnel). Des études complémentaires sur le déterminisme des points chauds, les facteurs génétiques associés et leur diversité seront nécessaires afin de pouvoir résoudre cette question.

Malgré toutes ces différences, la corrélation existante et relativement significative entre le taux de recombinaison familial et le taux de recombinaison populationnel permet de créer des cartes combinant les deux types de recombinaison et qui peuvent être utiles pour analyser de manière statistique des études génomiques. On peut ainsi citer un exemple avec une étude

récente de l'adaptation des moutons et des chèvres (Kim *et al.*, 2016). Dans cet article, un signal commun de « sélection » a été découvert en utilisant une statistique *iHS* (Integrated Haplotype Score), basée sur la comparaison de la fréquence d'une mutation donnée avec la longueur des haplotypes autour d'elle, (Voight *et al.*, 2006) dans les deux espèces. Or, cette signature coïncide parfaitement avec la région de très faible recombinaison que nous avons identifiée sur le chromosome 10. La statistique *iHS* a déjà été montrée comme étant très fortement influencée par la variation du taux de recombinaison, et conduit souvent à interpréter des régions de très faible recombinaison comme étant des signatures de sélection (O'Reilly *et al.*, 2008, Ferrer-Admetlla *et al.*, 2014). L'utilisation de cartes génétiques précises, comme celles que nous fournissons ici, peut donc aider à annoter et interpréter de tels signaux de sélection.

VI. Le biais d'usage des points chauds : autres pistes d'étude

Comme je l'ai indiqué dans le Chapitre 1, deux phénotypes de recombinaison sont communément étudiés : la variation inter-individuelle du taux de recombinaison et la variation inter-individuelle de la localisation de la recombinaison, ou biais d'usage des points chauds. S'il nous a effectivement été possible d'étudier le premier (voir Chapitre 3), nous n'avions cependant pas assez de données pour étudier le second. Il nous fallait soit plus d'animaux génotypés sur la puce 600K, soit des familles beaucoup plus grandes génotypées sur la puce 50K afin de cumuler suffisamment de méioses et d'avoir une très bonne résolution de la localisation des crossing-overs. Cependant, ces deux solutions sont trop coûteuses. Il aurait également pu être utile d'utiliser des approches directes, telles que le « sperm-typing », mais c'est une technique très lourde et coûteuse qui n'est pas adaptée pour une étude populationnelle.

Néanmoins, nous avons testé une autre approche statistique afin d'étudier ce phénotype.

VI. 1. *Essais d'une méthode indirecte pour détecter ce phénotype*

Afin d'étudier le deuxième phénotype, le biais d'usage des points chauds, nous avons tenté d'utiliser une méthode permettant d'estimer la variation d'intensité de recombinaison locale entre

les individus. Pour cela, nous sommes repartis du modèle de Poisson permettant d'estimer les taux de recombinaison (voir « Chapitre II. II. 2. ») et nous avons intégré un paramètre λ_{js} permettant de quantifier l'intensité de la recombinaison dans l'intervalle ***j*** et pour l'individu ***s***. Nous obtenons donc la formule d'estimation des taux de recombinaison suivante :

$$y_{sj} \mid c_j \sim \text{Poisson}[(0{,}01 * l_j * c_j * M_s * (R_s/R) * \lambda_{js}]$$

Nous avons ensuite essayé d'estimer les λ_{js} et d'utiliser leur variance sur l'ensemble des intervalles, pour chaque individu ***s***, comme phénotype de biais d'usage. Cependant, nous avons constaté que ce phénotype était extrêmement corrélé aux nombres de méioses observées, sauf peut-être pour les quelques individus ayant plus de 100 méioses. Nous en avons déduit que cette approche ne pouvait éventuellement fonctionner que si nous disposions de beaucoup d'individus ayant de nombreuses méioses.

VI. 2. *Recherche du gène PRDM9 à l'aide de motifs d'ADN spécifiques*

Cette étude a fait l'objet d'un stage de M1 en bio-informatique par Emilie Delpuech, que j'ai co-encadrée de Juin à Août 2016.

Le gène *PRDM9*, comme indiqué dans le Chapitre 1, est actuellement le seul gène connu de détermination de la localisation des points chauds chez certains Mammifères. La variation de *PRDM9* chez les Mammifères se joue surtout sur le nombre et la composition des doigts de zinc, la variation peut être inter-espèces, mais également intra-espèces. Les différents variants existants de *PRDM9* permettent à la protéine de reconnaître et de se fixer sur différents sites de l'*ADN* (Ahlawat *et al.*, 2016 a.). La rapide évolution des doigts de zinc de *PRDM9* montre de forts signaux de sélection positive et concertée à travers différentes espèces. Les doigts de zinc ont une forte homologie, excepté aux positions -1, 2, 3 et 6 qui déterminent la spécificité de fixation sur l'*ADN* et qui sont sélectionnées positivement (Pabo *et al.*, 2001, Oliver *et al.*, 2009).

PRDM9 possède une structure en doigt de zinc C2H2 ; structure qui possède deux cystéines et deux histidines, verrouillées par un atome de zinc divalent (**voir Figure 22**). La protéine est formée d'une hélice Alpha et d'un feuillet Béta aux brins antiparallèles. Ce sont les hélices qui portent les sites des doigts de zinc (Klug, 2010). Les doigts de zinc C2H2 sont le domaine de liaison à l'*ADN* le plus couramment trouvé chez les animaux et les plantes. Ils possèdent deux cystéines

conservées en partie N-Terminale et deux histidines conservées en C-Terminale. En plus de ces quatre acides aminés, chaque doigt de zinc peut contenir trois autres acides aminés, tels que la tyrosine, la phénylalanine ou la leucine, jouant un rôle dans la stabilité de la structure (Klug, 2010). Le contact entre le doigt de zinc et l'*ADN* s'effectue au niveau de l'hélice alpha et la liaison s'effectue par l'intermédiaire de liaisons hydrogènes spécifiques ou d'interactions hydrophobes. Les acides aminés de l'hélice, en position -1, 3 et 6, reconnaissent trois paires de bases consécutives sur le brin d'*ADN* (Klug, 2010).

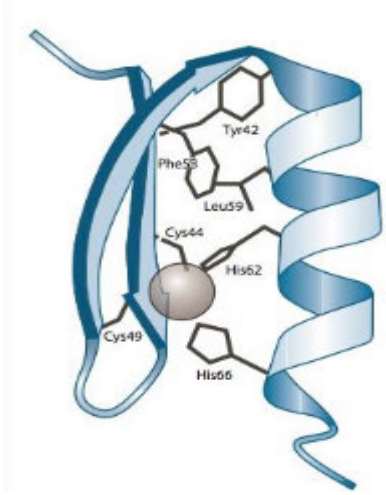

Figure 22 : Schéma d'un doigt de zinc (Klug, 2010)

Représentation d'un des doigts de zinc de la protéine *PRDM9* avec les deux cystéines, les deux histidines et la molécule de zinc au centre. Trois autres acides aminés sont également présents et jouent un rôle dans la stabilité de la protéine.

VI. 2. a. *Recherche de la séquence de PRDM9 chez le mouton*

Suite à diverses études, le gène *PRDM9* a été obtenu par amplification chez le bovin *Bos Taurus* (Ahlawat *et al.*, 2016 a.). Les amorces[17] données par cette étude (*F-TGCTCTCTGGCCTTCTCCAGTCAGAA* et *R-GCTGCAGTAATTCTCCTGTGAC*) ont été utilisées afin de récupérer les séquences du gène de la vache et du mouton dans les bases de données Ensembl et

17 L'amorce est une courte séquence d'*ARN* ou d'*ADN* complémentaire du début d'une matrice, servant de point de départ à la synthèse du brin complémentaire de cette dernière matrice par une *ADN* polymérase, notamment lors d'une *PCR*.

NCBI. Cependant, actuellement, seule la séquence du gène de la vache est disponible (Gene ID du *NCBI* : 100190914). La vache et le mouton étant deux espèces proches, il est possible d'utiliser les amorces bovines pour rechercher la séquence ovine, grâce à des méthodes de recherche de séquence par alignement, qui permettent de trouver des régions similaires entre deux ou plusieurs séquences. Pour cela, la technique du BlastN a été privilégiée dans un premier temps. Elle permet de comparer les amorces, qui sont des séquences nucléotidiques, avec une base de données regroupant des séquences nucléotidiques. Nous l'avons réalisée sur le site du *NCBI* où nous avons conservé les paramètres par défaut. Nous avons tout d'abord recherché les positions des amorces sur le génome ovin *Ovis aries*, puisque *PRDM9* y a été annoté, en alignant les amorces bovines sur le génome ovin. C'est la race ovine Texel qui a été séquencée et qui a permis d'obtenir la séquence de référence. Le BlastN retourne les occurrences de séquences similaires à celles données en entrée (ici, les amorces bovines), ainsi que leur position sur les chromosomes. Nous avons ensuite étudié la distribution de ces occurrences. Nous avons également refait ces analyses sur la base de données Ensembl, bien que le gène n'a pas encore été annoté.

Afin de savoir si les séquences sont proches, il faut étudier les valeurs de e-values données par les sorties du BlastN. Ces valeurs sont une statistique de test correspondant à l'espérance de pouvoir retrouver le score d'alignement, obtenu avec le BlastN, entre la séquence étudiée et la séquence de référence, uniquement grâce au hasard. Ainsi, plus ces valeurs sont faibles, voire nulles, plus il est improbable que le score d'alignement ait été obtenu par hasard et plus les séquences sont proches. On étudie également la longueur de l'alignement ; on recherche des séquences avec une taille d'alignement similaire. D'après une étude précédente, *PRDM9* est supposé se trouver sur le chromosome 1 du mouton (Ahlawat *et al.*, 2016 a.), cependant d'autres paralogues existent, sur les chromosomes X et 5 notamment. Les valeurs de e-value les plus faibles ont été effectivement obtenues sur le chromosome 1, entre les positions 275 609 926 et 275 609 951 paires de bases sur le brin « reverse » (brin « antisens », brin complémentaire au brin utilisé lors de la transcription de l'ADN) pour la première amorce et entre 275 608 940 et 275 608 959 paires de bases sur le « brin forward » (brin « sens) pour la deuxième. Les correspondances trouvées pour la deuxième amorce sur le chromosome 1 ne se placent cependant qu'en 6ème position, les plus faibles valeurs de e-value ayant été détectées pour les chromosomes 2 et 19. Cela montre que cette amorce n'est *a priori* pas spécifique du gène *PRDM9* ovin et que donc le

gène présente des divergences avec celui des bovins, puisque l'amorce a été designée spécialement pour cette espèce.

Avec la base de données Ensembl, nous avons obtenu une potentielle séquence pour le gène *PRDM9* sur le chromosome 1 ovin. Cependant, nous n'observons que 3 exons sur ce potentiel gène (**voir Figure 23**), or il a été décrit comme en possédant au moins 10 chez le mouton (Ahlawat *et al.*, 2016 a.).

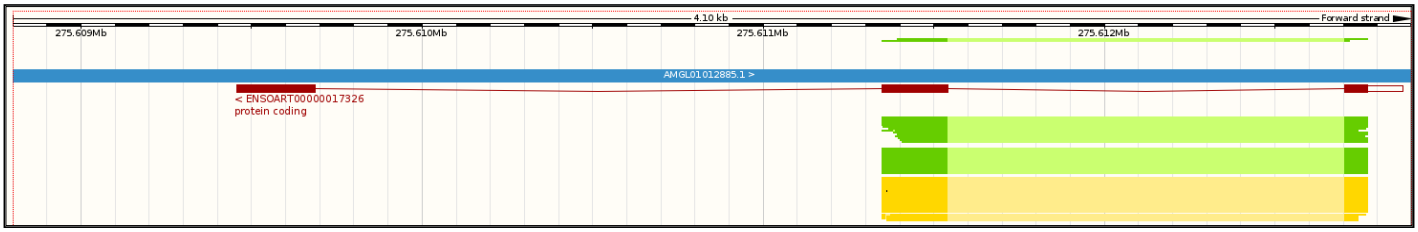

Figure 23 : Séquence de PRDM9 de Ovis aries issue de Ensembl

Observation des 3 exons (rectangles rouges) disponibles du gène *PRDM9* sur la base de données Ensembl. Le gène est compris entre les positions 275608800 pb et 275612895 pb sur le chromosome 1.

Néanmoins, sur ces 3 exons, un est porteur de doigts de zinc. Chez le mouton, c'est l'exon 10 du gène qui est porteur de 9 séquences en doigts de zinc, répétées successivement (Ahlawat *et al.*, 2016 a.). L'étude de notre exon ne révèle que 3 séquences en doigts de zinc. Il semble donc que le gène détecté soit tronqué. *PRDM9* se trouverait à l'extrémité du chromosome 1 du mouton. L'analyse de la séquence de la fin de ce chromosome montre que l'assemblage du génome se finit juste après les 3 exons obtenus ; le 3ème exon se terminant à 275 612 875 pb et le chromosome 1 ayant une taille de 275 612 895 pb. Entre ces deux positions, seule une suite de *N* est observable, indiquant que la séquence est inconnue. L'hypothèse la plus pertinente à l'issue de ces résultats est donc l'absence d'une partie du génome à l'extrémité C-Terminale du chromosome 1. Le génome a probablement été mal assemblé dans cette région, ce qui a conduit à tronquer le gène *PRDM9* et nous n'avons finalement accès qu'à la fin du gène. En effet, il se trouve sur le brin Reverse, donc présent de l'extrémité 3' vers 5', c'est pourquoi sur le premier exon que nous avons détecté nous observons des doigts de zinc, il s'agit en fait du dernier exon, l'exon 10. Nous avons vérifié le nombre de séquences répétées dans cette partie du génome grâce à une comparaison des séquences de l'exon 10, entre la vache (variation de 6 à 8 doigts de zinc (Ahlawat *et al.*, 2016

b.)), l'Homme (13 doigts de zinc pour l'allèle A (Myers *et al.*, 2010)) et le mouton (**voir Figure 24**), après avoir récupéré les séquences de *PRDM9* de l'Homme et la vache, à disposition dans les bases de données du *NCBI*. Pour cela, nous avons utilisé le logiciel Dotmatcher, disponible sur Emboss. Il permet de réaliser des « dotplots ». Ces graphiques prennent en entrée deux séquences et mettent notamment en évidence des séquences répétées lorsqu'elles sont comparées entre elles. Les paramètres par défaut ne permettant pas de mettre en avant les principales répétitions, nous avons choisi de paramétrer le logiciel avec une fenêtre de taille 50 ou 60 et un seuil de 60.

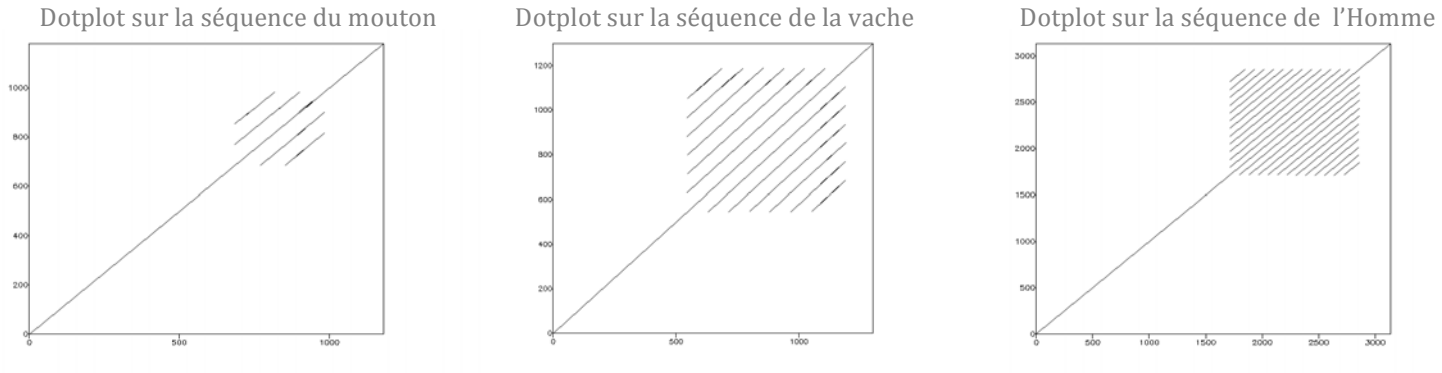

Figure 24 : Observation des séquences de l'exon 10 du mouton, de la vache et de l'Homme

Réalisation de « dotplots » permettant de repérer les séquences répétées chez le mouton, chez la vache et chez l'Homme. Les séquences répétées correspondent ainsi aux différents doigts de zinc dans ces espèces.

Chez le mouton, seulement trois séquences en doigts de zinc sont retrouvées, alors que nous voyons 7 répétitions chez la vache et 13 chez l'Homme. Ces résultats étaient notre hypothèse d'une absence de *PRDM9* dans son intégralité sur le génome du mouton. Les sorties du BlastN pour Ensembl et le *NCBI* révèlent des occurrences toujours dans les premières, étant donné leur e-values proches de 0. Ces occurrences sont retrouvées sur un « scaffold » non annoté sur le génome.

VI. 2. b. *Le gène PRDM9 est contenu dans un « scaffold »*

Un « scaffold » est composé de « contigs »[18], mais il n'est pas assemblé au sein de la séquence de référence du génome. C'est donc un morceau de séquence présentant des différences par rapport à un génome de référence et qui ne peut donc pas être assemblé. Le « scaffold » découvert, identifiant *JH922946*, présente une très forte ressemblance avec la séquence de *PRDM9* recherchée (**voir Figures 25 et 26**).

Figure 25 : « Scaffold » JH922946 de Ovis aries avec 10 exons annotés

Récupération d'un « scaffold » annoté sur la base de données Ensembl comprenant les 10 exons annotés du gène *PRDM9*.

18 Un contig est une séquence obtenue à la suite de l'assemblage de séquences présentant des zones identiques et chevauchantes à leurs extrémités 3' et 5'.

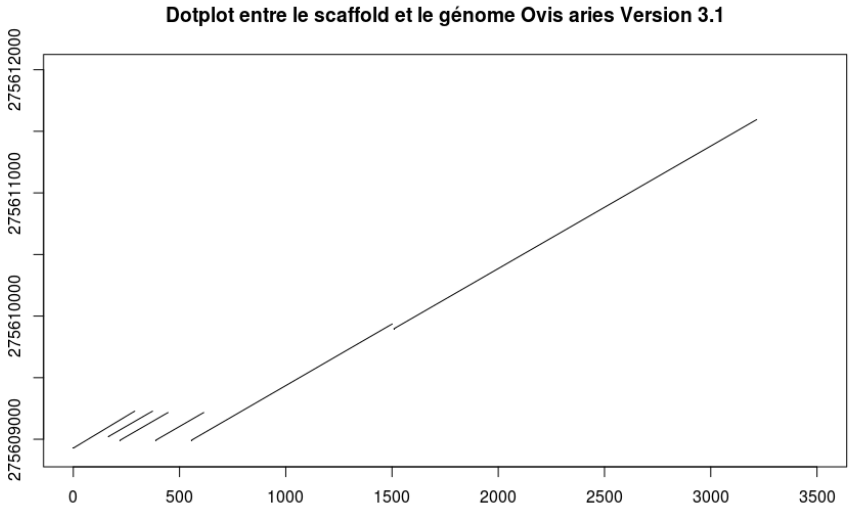

Figure 26 : Comparaison entre le « scaffold » et la fin du chromosome 1

On remarque une très grande ressemblance entre les séquences du « scaffold » et la fin du chromosome 1 du génome de *Ovis aries*, ce qui peut suggérer que ce « scaffold » devrait être assemblé à la fin du génome.

Sur cette séquence, un total de 10 exons a été détecté et annotés. Il semble donc bien s'agir des 10 exons annoncés par Ahlawat *et al.* (2016 a.) dans leur étude. Grâce aux amorces, nous avons pu retrouver l'exon 10 situé entre les bases 20 et 1216, donc au début de la séquence du « scaffold », sur le brin Reverse.

L'alignement entre ce « scaffold » et le génome ovin révèle que 2 700 pb du génome s'alignent avec le « scaffold » entre les positions 275 608 000 pb et 275 612 000 pb, partie C-Terminale du chromosome 1 ovin. En revanche, les positions correspondantes sur le « scaffold » sont en partie N-Terminale, entre 1 et 3217 pb. Des BlastsN entre notre « scaffold » et les séquences de *PRDM9* de l'Homme, de la vache, du cheval et de la chèvre ont permis de montrer la très grande ressemblance de notre séquence avec les séquences des gènes orthologues. Avec la vache, le premier résultat obtenu correspond précisément au gène *PRDM9* localisé sur le chromosome 1 bovin.

Tous ces résultats confirment donc l'hypothèse précédente : la séquence du chromosome 1 de *Ovis aries* est incomplète et le gène *PRDM9* ne semble donc pas avoir pu être annoté. Cependant, nos analyses tendent à positionner le gène *PRDM9* sur le « scaffold » *JH922946*. Il se pourrait qu'il n'ait pu être ajouté à la partie C-Terminale de la séquence du chromosome 1 en

raison d'une différence de séquences, potentiellement due à des insertions ou délétions. De plus, les extrémités des chromosomes comportent souvent des erreurs et sont donc souvent tronquées.

VI. 2. c. *Etude de la structure du gène PRDM9*

Le gène *PRDM9* est découpé en plusieurs exons et introns[19]. Chez la vache, 10 exons ont été trouvés (Ahlawat *et al.*, 2016 b.), parmi lesquels seul le dernier présente des séquences en doigts de zinc de type C2H2. Les amorces présentées précédemment nous ont permis de rechercher l'exon de *PRDM9* contenant les doigts de zinc. Chez les espèces où le gène a été étudié, cet exon a souvent une taille proche de 1 Kb. Le premier doigt de zinc est très souvent conservé entre les espèces, de ce fait il est exclu des analyses.

Chez la vache, les doigts de zinc sont des petites séquences de 23 acides aminés conservées et répétées en tandem dans le génome, en particulier dans la séquence de l'exon 10 (Ahlawat *et al.*, 2016 b.).

VI. 2. c. a. Obtention de la séquence protéique de *PRDM9*

Chez le mouton, la taille du « scaffold » découvert, *JH922946*, est de 16,283 pb. Les 10 exons ont été retrouvés sur ce « scaffold ». L'exon 10 est le plus grand, c'est lui qui contient les séquences en doigts de zinc. Il a une taille de 500 pb et se situe dans les 1 400 premières bases du « scaffold ».

Etant donné que nous avions finalement confirmé la présence de *PRDM9* dans la séquence nucléotidique, il nous fallait ensuite connaître sa séquence protéique. Dans les bases de données, notamment Ensembl, sur laquelle nous avions pu récupérer la séquence nucléotidique, une séquence protéique du gène était effectivement disponible. Cependant, elle ne correspondait qu'à la traduction des premiers exons du « scaffold », et nous n'avions donc pas à disposition la séquence protéique de l'exon 10. Nous l'avons donc obtenue grâce à un programme informatique écrit avec le package « BioPython » qui permet de réaliser la traduction d'une séquence

19 Un intron est une portion d'un gène qui est transcrite en *ARN* et qui est ensuite éliminée et donc non retrouvée dans l'*ARN* messager.

nucléotidique. Le programme renvoie la séquence protéique de la séquence donnée en entrée. Pour tout codon-stop traduit, un astérisque est inséré dans la séquence protéique. Afin d'être sûr d'avoir l'exon 10 nous traduisons 3 000 paires de bases, ce qui nous donne une séquence finale protéique de 1 000 acides aminés.

Une fois les séquences protéiques obtenues, nous avons recherché les doigts de zinc. Nous avons pour cela réalisé des alignements multiples entre les séquences protéiques du mouton, de la vache, de l'Homme et de la chèvre (**voir Figure 27**).

```
QQRYSDPHSPSDKPEGREAKDRPQPLLKSIRLKRISRASSYSPRGQMRASGVHERITEEP
LFDYAFIQNSALRPHSVT-HT------REITLECRVCGKTFSKNSNLRR---H-------
----------------------------------------AFSKSSNLRR---H-------
QRRITFIQGSAHRQHNNT-QT------GEKTFECHVCRKAFGKSSNLRR---H-------
                                          ::.  .::*    *

STSQKLNPEDTGKLLMGAGVSGIIKVKYGECGQGSKDRSSLITNQRTHTGEKPYVCRECG
-------------EM---IHTGEKPHGCHLCGKAFTHCSDLRKHERTHTGEKPYGCHLCG
-------------EM---IHTGVKPHGCHLCGKSFTHCSDLRKHERIHTGEKLYGCHLCG
-------------EM---IHTGVKPHRCHLCGKSFTHCSDLRKHERIHTGEKLYGCHLCG
              :     :*        **:. .. *.* ..:* ***** * *: **

RSFSDKSNLITHQRTHTGEKPYVCRECGRSFSVKSHLITHQRTHTGEKPYVCRECGRSFS
KAFSKSSNLRRHEMIHTREKAQICHLCGKAFTHCSDLRKHERTHLGDKPYGCLLCGKAFS
KAFSKSYNLRRHEVIHTKEKPNECHLCGKAFAHCSDLRKHERTHFGEKPYGCHLCGKTFS
KAFSKSYNLRRHEVIHTREKPNECHLCGKAFVHCSDLKQHERTHFGEKPYGCHLCGKTFS
::**.. **  *:  ** **   *: **::*   *.*  *:*** *:*** *  **::**

VKSHLITHQRTHTGEKPYVCRECGRSFSVKSHLITHQRTHTGEKPYVCRECGRSFSDKSN
KCSYLRQHERTHNGEKPYECHLCGKAFSHCSHLRQHERSHNGEKPHGCHLCGKAFTESSV
KTSYLRQHERTHNGEKPYGCHLCGKAFTHCSHLRKHERTHTGEKPYECHLCGKAFTESSV
KASYLRQHERTHNGEKPYGCHLCGKAFTHCSHLRKHERTHTGEKPYECHLCGKAFTESSV
  *:*  *:***.***** *: **::*:  ***  *:*:*.****: *: **::*::.*

LITHQRTHTGEKPYVCRECGRSFSDKSNLITHQRTHTGEKPYVCRECGRSFSDKSNLITH
LKRHERIHTGEKPYECHVCGKAFTESSDLRRHERTHTGEKPYECHLCGKAFNHSSVLRRH
LRRHERTHTGEKPYECHLCWKAFTDSSVLKRHERTHTGEKPYECHLCGKTFNHSSVLRRH
LRRHERTHTGEKPYECHLCWKAFTDSSVLKRHERTHTGEKPYECHLCGKTFNHSSVLRRH
*  *:* ******* *: * ::*::.* *  *:********* *: **::*...* *  *

QRTHTGMCAGSV---GKAS-VLSQILSHTRGHTREKPYVCRECEWIMTNK----------
ERTHTGEKPYECNICGKAFNRSYNFRLHRRVHTGEKPYVCPLCGKAFSKFFNLRQHERTH
ERTHTGEKPYECNICGKAFNRSYNFRLHKRIHTGEKPYKCYLCGKAFSKYFNLRQHENSC
ERTHTGEKPYECNICGKAFNRSYNFRLHKRIHTGEKPYKCYLCGKAFSKYFNLRQHERTH
:*****    .    ***     ::  * * ** **** *  *   :::

-------------------
TKKAMNM------------
YKGNK--------------
AVKVINEDSPPIPSKYKHS
```

Figure 27 : Alignement multiple de l'exon 10 de PRDM9 entre le mouton, la vache, l'Homme et la chèvre

Utilisation de Clustal Omega pour réaliser des alignements multiples des séquences protéiques de *PRDM9* entre différents Mammifères. La première ligne correspond à la séquence protéique du mouton, la deuxième à celle de la vache, la troisième à celle de l'Homme et la dernière est la séquence protéique de la chèvre. Un très bon alignement des séquences est observé, d'où une grande conservation du gène chez les Mammifères.

Les séquences protéiques de *PRDM9* pour ces espèces sont à disposition sur les bases de données, comme celle du *NCBI*. Les séquences s'alignent extrêmement bien entre elles, malgré la grande

différence phylogénétique entre les espèces, notamment entre l'Homme et le mouton. Cela montre bien que, malgré l'évolution rapide de *PRDM9*, des parties du gène sont très conservées entre les Mammifères, notamment au niveau des deux cystéines et des deux histidines. Entre ces acides aminés, de la variabilité est observée.

VI. 2. c. b. Etude des doigts de zinc de *PRDM9* ovin

Afin d'évaluer la ressemblance de notre séquence protéique de *PRDM9* avec celles existantes dans d'autres espèces, nous avons réalisé de nouveaux Blasts. Nous constatons une plus grande ressemblance avec les séquences de la vache et de l'Homme. Ce qui est en faveur d'une bonne traduction de notre séquence nucléotidique et de la présence du gène *PRDM9* sur le « scaffold ».

Nous avons ensuite cherché à caractériser les doigts de zinc de notre exon 10. Nous avons donc recherché une séquence d'environ 20 acides aminés débutant par deux cystéines, séparées par deux acides aminés, et se terminant par deux histidines, séparées par trois acides aminés quelconques. Grâce à notre séquence protéique, nous avons détecté 7 doigts de zinc distincts, ainsi que le 1er doigt de zinc qui est retrouvé chez quasiment toutes les espèces et donc conservé au sein de *PRDM9*. Etant donné sa forte conservation, ce doigt de zinc est donc écarté des analyses suivantes. Mis à part ce 1er doigt de zinc et les 7 autres répétés en tandem, un 9ème doigt de zinc a été observé, cependant, il ne possède qu'une seule des deux premières cystéines et une seule histidine en fin de séquence. Malgré ces différences, la séquence de ce doigt de zinc est identique aux 7 autres et peut donc être prise en compte. Ce doigt de zinc se trouve au début du lot de répétitions et est séparé du 2ème doigt de zinc par un motif spécifique, identifié comme un séparateur des séquences répétées en doigts de zinc. Pour confirmer nos résultats, nous avons utilisé un logiciel en ligne, le *DNA* sequence Logo Generator, disponible à l'adresse suivante : http://zf.princeton.edu/, qui détermine également 8 doigts de zinc issus de nos séquences nucléotidiques et protéiques.

Puisque le programme utilisé pour la traduction de la séquence nucléotidique souligne les codons-stops à l'aide d'un astérisque, nous avons pu facilement identifier le début et la fin de notre séquence protéique et donc s'assurer de la présence de tous les doigts de zinc.

Nous avons ensuite caractérisé la variabilité de nos doigts de zinc (**voir Tableau 5**). Pour cela,

nous avons aligné les 9 séquences des différents doigts de zinc décrits précédemment et nous avons recherché de potentielles variations au niveau des acides aminés. D'après la littérature, les sites soumis à la sélection chez le mouton se trouvent majoritairement en position -5, -1, 3 et 6 sur les doigts de zinc (Ahlawat *et al.*, 2016 a.). Dans les séquences de nos doigts de zinc, nous n'observons pas de variabilité en position 6, en revanche les autres positions montrent une relative variation. Les allèles déterminés sont similaires à ceux déjà décrits chez le mouton (Ahlawat *et al.*, 2016 a.).

Tableau 5 : Diversité des séquences des doigts de zinc dans la séquence protéique de l'exon 10 de PRDM9 présent sur le « scaffold » JH92946 chez le mouton

					-5				**-1**			**3**			**6**					
Y	G	E	C	G	Q	G	S	K	D	R	S	S	L	I	T	N	Q	R	T	H
C	R	E	C	G	R	S	F	S	D	K	S	N	L	I	T	H	Q	R	T	H
C	R	E	C	G	R	S	F	S	V	K	S	H	L	I	T	H	Q	R	T	H
C	R	E	C	G	R	S	F	S	V	K	S	H	L	I	T	H	Q	R	T	H
C	R	E	C	G	R	S	F	S	V	K	S	H	L	I	T	H	Q	R	T	H
C	R	E	C	G	R	S	F	S	D	K	S	N	L	I	T	H	Q	R	T	H
C	R	E	C	G	R	S	F	S	D	K	S	N	L	I	T	H	Q	R	T	H
C	R	E	C	G	R	S	F	S	D	K	S	N	L	I	T	H	Q	R	T	H

Les positions sont indiquées relativement au début de l'hélice alpha de chaque doigt de zinc. Chaque lettre correspond à un acide aminé. Une variabilité au niveau des positions -5, -1 et 3 est remarquée, mais pas au niveau de la position 6. Les lettres correspondent aux différents acides aminés ; C : Cystéine, D : Aspartate, E : Glutamate, F : Phénylalanine, G : Glycine, H : Histidine, I : Isoleucine, K : Lysine, L : Leucine, N : Asparagine, Q : Glutamine, R : Arginine, S : Sérine, T : Thréonine, V : Valine, Y : Tyrosine.

VI. 2. d. *Recherche de la séquence de PRDM9 dans les points chauds de recombinaison*

VI. 2. d. a. Prédiction de motifs spécifiques à l'aide d'une séquence consensus

Puisque qu'il ne nous est pas possible de rechercher le gène *PRDM9* par une méthode de

détection de *QTL* sur un phénotype particulier, ici le biais d'usage des points chauds, nous avons essayé de détecter ce gène par une méthode indirecte. Nous avons ainsi cherché à prédire les séquences d'*ADN* ovines sur lesquelles les doigts de zinc de *PRDM9* viennent se fixer. Pour ce faire, grâce au logiciel en libre service présenté précédemment, nous avons généré un motif *ADN*, représentant les sites de fixation des doigts de zinc, et une matrice de poids position (*PWM*) à partir de la séquence protéique. Une *PWM* dérive d'un ensemble d'alignements multiples et est un outil important pour la recherche de motifs. Cette approche utilise une *SVM* (Support Vector Machine), un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination, pour prédire la liaison de l'*ADN* avec les doigts de zinc. Le motif et la matrice sont calculés à partir du nombre de doigts de zinc que l'on aura sélectionné. A partir de cette *PWM* nous pouvons déduire une séquence consensus à partir de laquelle nous retrouvons les sites de liaison à l'*ADN* des doigts de zinc. Cette séquence consensus est estimée à partir des scores de probabilité de chaque base à chacune des positions, ainsi qu'à partir d'un score d'entropie, qui reflète la qualité de l'alignement ; plus le score est faible, plus l'alignement est correct. Nous ne gardons que les nucléotides pour lesquels l'entropie est faible, pour les autres, nous écrivons un N, signifiant que ce peut être n'importe quelle base.

La séquence consensus estimée à partir des scores de probabilité est ainsi composée des 25 nucléotides suivants : *TACTACTACTGTTGTTGTTACTTTG* (**voir Figure 28**).

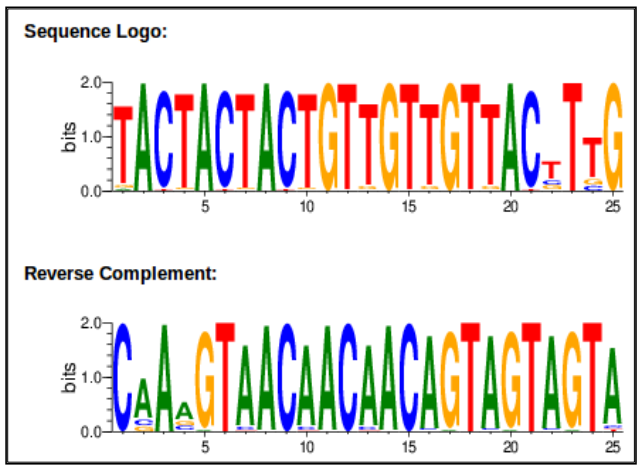

Figure 28 : Logo obtenu avec la séquence protéique et la détection des doigts de zinc

Observation du logo obtenu avec nos doigts de zinc. Le logo est très riche en bases *A* et *T*, mais très pauvre en bases *C* et *G*.

Pour affiner la sélection, il faut également prendre en compte le score d'entropie ; plus ce score est faible, plus la base ayant le plus grand score de probabilité à cette position a de chance d'être confirmée. Ainsi, nous ne gardons que les bases avec un score d'entropie inférieur à 0,1, pour les autres bases, nous mettons un N pour indiquer que n'importe quel nucléotide peut être présent à cette position. Suite à l'application de ce critère, nous obtenons le motif suivant, contenant 12 bases stables : *NTNNTNNTNNTANCANCANTNNANC*. Comparé au motif de 13 nucléotides découvert par Myers *et al.* (2008), notre motif est pauvre en bases *C* et *G*, or les points chauds sont des séquences du génome où le taux de *GC* est plus important que dans le reste du génome. Au contraire, les bases stables de notre motif sont principalement des *A* et des *T*, *a priori* plutôt enrichies dans les points froids.

Malgré tout, nous avons cherché à savoir si notre séquence consensus pouvait se retrouver dans nos points chauds. Pour cela, nous avons utilisé les 20 000 points chauds obtenus avec un *FDR* de 0,1%, seuil très stringent permettant d'être plus sûr de nos résultats. Dans ces 20 000 points chauds, grâce à un programme Python 2.7 obtenu avec le package « Numpy », nous avons extrait toutes les séquences de la même taille que notre motif et nous avons calculé leurs scores de probabilité et d'entropie. Nous avons fait de même avec les points froids, qui permettent de mettre en évidence d'éventuelles différences avec les points chauds. Les points froids ont été définis comme des séquences non identifiées en points chauds et qui ont une taille physique identique à celle des points chauds étudiés. Les scores ont ensuite été analysés sous R afin de voir si certains motifs étaient différents entre points chauds et points froids, voire spécifiques des points chauds. Nous avons analysé la distribution des scores et nous l'avons comparée à la loi Normale, nous avons également comparé les distributions des motifs des points chauds et des points froids à l'aide de diagrammes quantile-quantile. Ces graphes permettent de comparer deux distributions entre elles et un alignement sur la diagonale indique la présence d'une identité de loi.

A la suite de ces analyses, nous remarquons que parmi tous les motifs extraits des points chauds, aucun ne présente toutes les bases stables de notre motif. Nous nous sommes donc demandés s'il ne valait pas mieux supprimer les doigts de zinc incomplets lors de la création du logo. Lorsque nous ne prenons en compte que les 7 doigts de zinc complets, la séquence suivante est obtenue (pour le critère des scores de probabilité) : *TACTACTACTGTTGTTGTTACT*, ce qui donne

après contrôle sur les scores d'entropie, *NTNNTNNTNNTANCANCANTNN*. Là-encore, le motif est très pauvre en *CG* et aucun de nos points chauds ne possède les bases prédites. De plus, aucune différence significative n'a pu être détectée entre les points chauds et les points froids ; les scores suivant la même distribution entre les deux types de séquence.

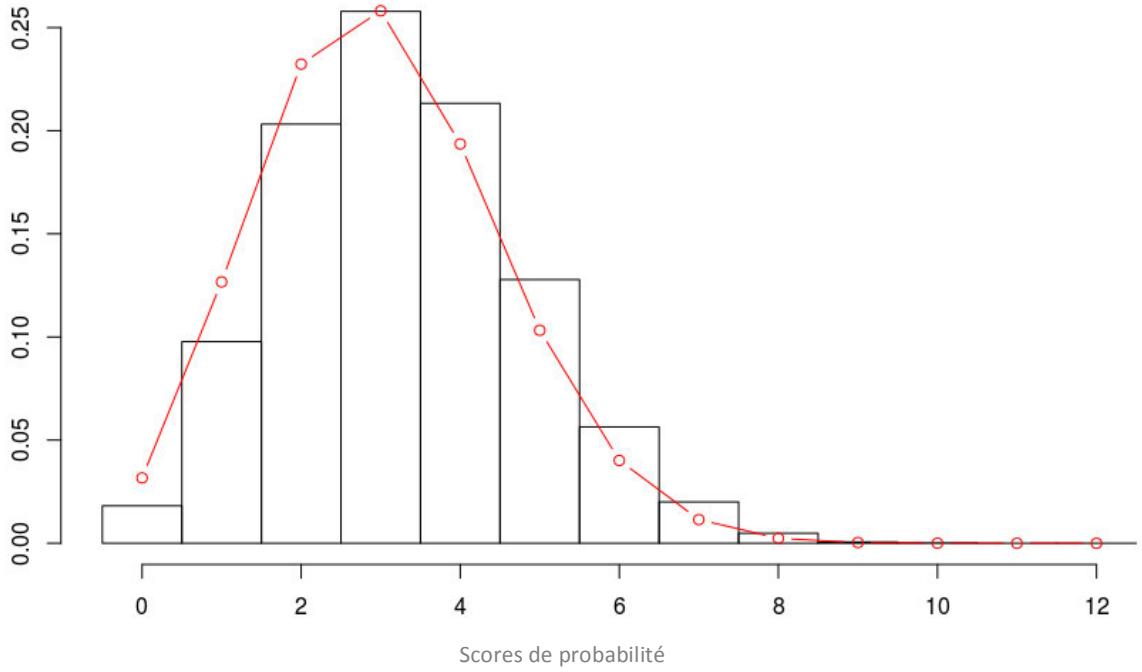

Figure 29 : Histogramme représentant la distribution des scores des motifs

Distribution des scores de présence du motif prédit (histogramme) comparée à la distribution attendue du score pour un motif aléatoire (ligne rouge).

La comparaison entre la distribution des scores des motifs des points chauds et la distribution attendue des ces scores, (**voir Figure 29**), indique que les motifs sont répartis de manière plutôt aléatoire dans les points chauds et les points froids. Quant à l'étude des diagrammes quantile-quantile, elle révèle qu'il est impossible de différencier les points chauds et les points froids sur la base des motifs, étant donné qu'une droite est obtenue (**voir Figure 30**).

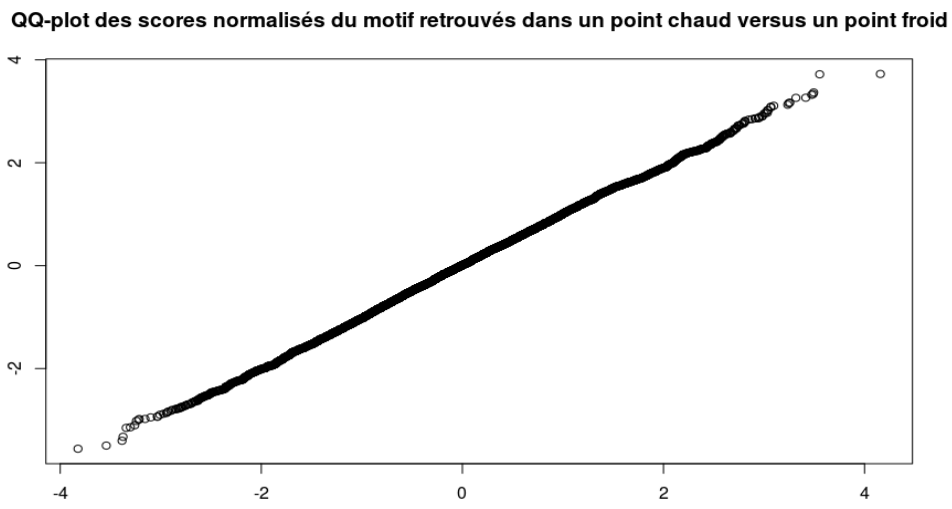

Figure 30 : Diagramme quantile-quantile des scores normalisés obtenus en comparant un point chaud et un point froid

Comparaison des scores entre un point froid en ordonnée et un point chaud en abscisse. Aucune différence ne peut être notée étant donné qu'une droite est obtenue.

Il n'est donc pas possible de prédire des motifs spécifiques des points chauds à l'aide de notre séquence consensus. Cela peut être dû au fait que cette dernière ait pu être mal prédite par la matrice *PWM*, ou bien du fait que de potentielles erreurs dans notre séquence protéique engendrent une fausse *PWM*. Ce qui expliquerait alors la pauvreté de notre séquence consensus en bases *CG*.

VI. 2. d. b. Recherche de séquences répétées dans les points chauds

Etant donné que nous n'avons pas pu prédire de séquences spécifiques aux points chauds à l'aide de notre séquence consensus, nous avons décidé d'étudier directement les séquences des points chauds et d'y rechercher un enrichissement en motifs spécifiques. Comme précédemment, nous allons comparer les séquences obtenues dans les points chauds avec celles des points froids. Pour chacun des 23 198 points chauds détectés, nous avons déterminé la taille physique et le taux de *GC*. Ces deux variables sont normalisées entre 0 et 1, grâce à la fonction « ecdf » de R qui permet de représenter la fonction de répartition empirique de la taille et du taux de *GC*. Nous

avons associé à chaque point chaud un point froid possédant des caractéristiques similaires ; c'est-à-dire une taille et un taux de *GC* semblables. Ceci permet d'obtenir 23 198 points froids, certains points chauds partageant le même point froid. Par la suite, nous avons établi d'autres critères de sélection, en ne conservant notamment que les points froids se trouvant sur le même chromosome que le point chaud, à moins de 60 Kb de lui, sachant que les points chauds sont séparés d'une distance d'environ 56 Kb. De plus, nous ne conservons que des points chauds avec un log10(λ) supérieur à 1 ; λ représentant l'intensité de recombinaison du point chaud, afin d'être certain de nos points chauds. Nous avons ainsi obtenu 22 680 points chauds et autant de points froids.

Sur cette sélection de points chauds et de points froids, nous avons donc recherché un potentiel enrichissement en motifs spécifiques. Pour cela, nous avons utilisé le logiciel MEME. Ce logiciel permet de trouver des motifs complets dans des séquences. La taille des motifs désirés peut être définie par l'utilisateur, en indiquant une taille minimum et une taille maximum. Un motif est défini comme une petite séquence qui se répète plusieurs fois et MEME représente les motifs par des matrices de position ; chaque nucléotide ayant une certaine probabilité de se trouver à une position donnée, ce qui révèle un logo. Le logiciel prend en entrée un groupe de séquences et émet autant de motifs que l'utilisateur le souhaite. Des techniques de modélisation statistique sont utilisées pour choisir automatiquement la meilleure taille, le meilleur nombre d'occurrences, ainsi que la meilleure description de chaque motif. Il est également possible de donner au logiciel un deuxième jeu de séquences, servant de contrôle, et permettant de ne trouver que les motifs enrichis dans le premier jeu de séquences ; il s'agit d'un mode de recherche dit « discriminant ».

Nous avons extrait les séquences nucléotidiques de chacun des points chauds et des points froids associés, celles des points froids servant de contrôle et permettant de vérifier que les motifs détectés dans les points chauds ne sont pas enrichis dans les points froids. Nous avons tout d'abord essayé de retrouver notre séquence consensus, cependant, comme précédemment, aucun résultat probant n'a été obtenu. Dans un deuxième temps, nous avons utilisé les motifs disponibles dans MEME pour rechercher un ou plusieurs motifs caractéristiques des points chauds de recombinaison dans nos points chauds. Cependant, le fichier de séquences de nos 23 198 points chauds est trop important pour MEME, le seuil maximum étant de 1 000 séquences. Pour

ce qui est du mode de recherche discriminant, il n'a pas été possible de le mettre en place, car le fichier contrôle, avec les séquences des points froids, comporte plusieurs séquences dupliquées, c'est-à-dire que plusieurs points chauds ont le même point froid associé. Ces erreurs empêchent l'utilisation du logiciel MEME et donc la recherche d'un enrichissement en motifs spécifiques. Il aurait été possible de passer outre ces erreurs en supprimant les séquences répétées dans le jeu de données des points froids, ainsi que dans celui des points chauds associés, pour pouvoir obtenir un motif à l'aide de l'outil de prédiction du logiciel MEME.

VI. 3. *Conclusion intermédiaire : la recherche de PRDM9*

Puisqu'il n'était pas possible de rechercher *PRDM9* par des méthodes de détection de *QTLs*, nous l'avons recherché par des études indirectes. Nous avons ainsi confirmé sa présence sur le chromosome 1 du mouton, cependant, il est contenu dans un « scaffold » non assemblé qui devrait se placer en fin du chromosome 1. Dix exons ont été mis en évidence, avec le dernier comportant 9 doigts de zinc, dont 1 très conservé et 7 complets. A partir de ces doigts de zinc, une séquence consensus a pu être créée, cependant, elle est riche en *AT* et pauvre en *CG*, ce qui est en désaccord avec le motif majoritaire présent dans les points chauds humains. De plus, elle n'a pas été retrouvée dans nos points chauds, et, plus généralement, aucun motif particulier n'a été montré comme enrichi dans nos points chauds. Cette absence de résultats peut s'expliquer par une complexité de prédiction des motifs. En effet, les bases de données utilisées ne sont pas forcément spécifiques des moutons. Par ailleurs, l'utilisation d'une matrice *PWM* peut entraîner des mauvaises prédictions et c'est une méthode compliquée à mettre en œuvre. De plus, nos points chauds sont relativement grands (autour de 5 Kb), ce qui entraîne donc de grandes séquences et donc énormément de données qu'il n'est pas possible de traiter par méthodes informatiques. Il peut également y avoir des problèmes d'assemblage au niveau des points chauds, qui empêchent les logiciels de créer des motifs complets. De plus, la séquence de référence étant obtenue à partir de la race Texel, il est possible que la séquence de *PRDM9*, ses doigts de zinc ou les sites de sélection soient différents de ceux de la Lacaune. Il serait donc intéressant de générer les allèles spécifiques à la Lacaune de *PRDM9* par re-séquençage de la région génomique.

Chapitre 3 : Etude du Déterminisme Génétique

Chapitre 3 : Etude du déterminisme génétique de la variation inter-individuelle du taux de recombinaison

I. Observation du phénotype « taux de recombinaison individuel »

En utilisant les crossing-overs découverts avec notre jeu de données familial, nous avons cherché à savoir si les Lacaune, à l'instar des autres animaux d'élevage, présentaient une variation inter-individuelle du taux de recombinaison. Pour cela, nous avons estimé pour chacun des 345 pères leur taux de recombinaison sur l'ensemble du génome à partir de leur nombre observé de crossing-overs par méiose. A l'aide d'un modèle linéaire mixte, nous avons corrigé cette valeur pour différentes covariables : l'année de naissance de chacun des 345 pères, allant de 1997 à 2010, (covariable considérée comme un cofacteur avec 14 niveaux) et le mois d'insémination de la mère des descendants par un des 345 pères (cofacteur de 7 niveaux allant de Février à Août). La contribution des effets génétiques additifs a été estimée en incluant un effet père aléatoire avec une structure de covariance proportionnelle aux coefficients de la matrice de parenté. Ces derniers sont calculés à partir des informations de pédigrées.

Le modèle mixte peut donc s'écrire sous la forme suivante :

$$\mathbf{y_{so} = \mu + x_{so}\beta + u_s + e_{so}} \text{ avec :}$$

- $\mathbf{y_{so}}$: nombre de crossing-overs dans la méiose entre le père ***s*** et son descendant ***o***.
- $\boldsymbol{\mu}$: taux de recombinaison moyen dans la population après correction pour l'effet père.
- $\boldsymbol{\beta}$: vecteur des effets fixes (effets environnementaux).
- $\mathbf{u_s}$: déviation par rapport à la moyenne propre à chaque père ***s***. $\boldsymbol{u_s}$, suit une loi Normale : $\mathbf{u_s \sim N(0, A\sigma_s^2)}$.
- $\mathbf{e_{so}}$: effets aléatoires qui suivent une loi Normale : $\mathbf{e_{so} \sim N(0, I\sigma_e^2)}$.

- **A** : matrice de parenté entre les pères et $\boldsymbol{x_{so}}$, la matrice d'incidence des effets fixes.
- **I** : matrice identité.

Nous utilisons le logiciel BLUPf90 (Misztal *et al*., 2002), qui permet d'estimer les composantes de la variance du modèle en utilisant un *REML* (Restricted Maximum Likelihood) et nous récupérons les composantes des variances σ_e^2 et σ_s^2, qui permettent d'estimer l'héritabilité du phénotype. L'héritabilité, h^2, est calculée selon la formule suivante :

$$h^2 = \sigma_s^2/(\sigma_e^2 + \sigma_s^2)$$

Les valeurs génétiques u_s sont aussi estimées pour chaque père. Notre phénotype est donc considéré comme un index dans la suite des analyses et sur la **Figure 31**, nous observons qu'il varie selon les individus, ce qui est cohérent avec ce qui est observé dans les autres espèces.

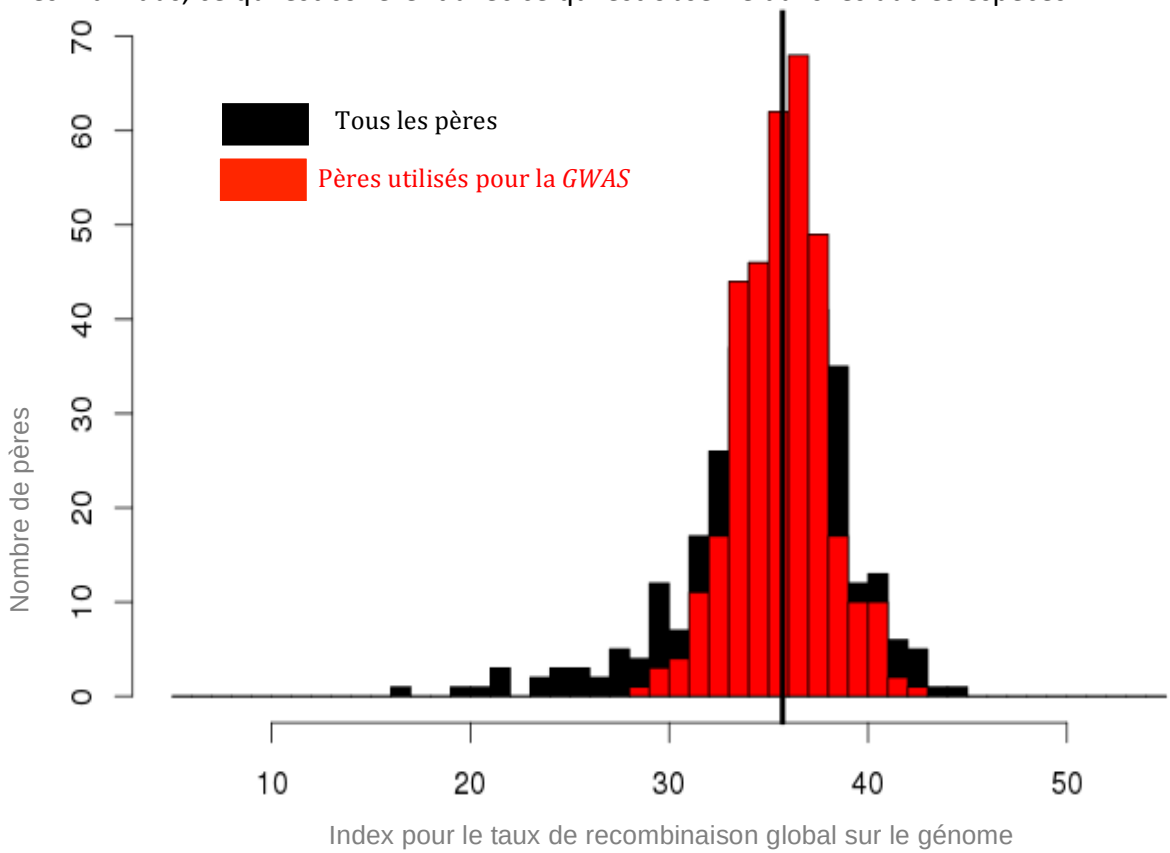

Figure 31 : Variation individuelle du taux de recombinaison parmi les Lacaune mâles

Distribution des valeurs génétiques additives du taux de recombinaison pour tous les pères Lacaune de notre jeu de données (en noir) et pour les 345 pères Lacaune utilisés dans nos analyses (en rouge). La ligne verticale noire correspond à la moyenne.

Aucune différence significative n'a été mise en évidence entre les années de naissance des pères, en revanche le mois d'insémination était significatif (p-valeurs de $1,3*10^{-3}$) (**Voir Figure 32**).

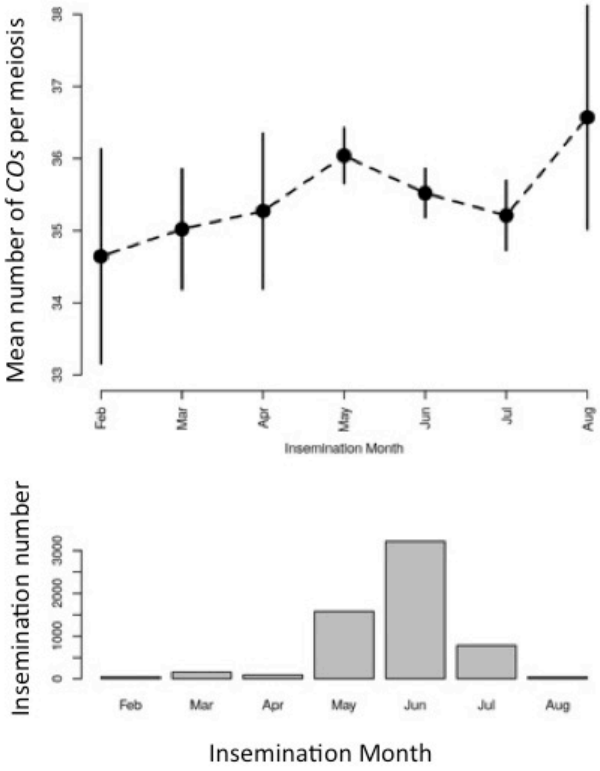

Figure 32 : Effet du mois d'insémination sur le nombre moyen de crossing-overs par méiose.

Graphe du haut : nombre moyen de crossing-overs par méiose pour chaque mois et intervalles de confiance à 95% (lignes verticales). Graphe du bas : nombre d'inséminations par mois

Nous avons ainsi remarqué que la moyenne, sur la population, du nombre de crossing-overs augmentait entre Février et Mai (passant de 34,92 crossing-overs par méiose à 36,27) pour ensuite rediminuer jusqu'à 35,46 crossing-overs par méiose en Juillet (la moyenne du mois d'Août n'est pas pertinente du fait d'un trop faible nombre d'observations). Il se pourrait donc que la recombinaison soit sensible à la température extérieure avec un optimum pour le mois de Mai. A partir des composantes des variances (**voir Tableau 6**), nous avons pu estimer l'héritabilité du taux de recombinaison individuel à 0,23.

Tableau 6 : Décomposition de la variation du taux de recombinaison individuel chez les mâles Lacaune

Phénotype	Nombre de pères	Variance génétique	Variance phénotypique	Héritabilité
Nombre de crossing-overs	345	6,86 (0,75)	29,73 (0,84)	0,23 (0,02)

Variance génétique, variance phénotypique et héritabilité du taux de recombinaison individuel pour 345 pères. Les chiffres entre parenthèses correspondent aux erreurs d'estimation.

II. La détection de *QTLs* suite à l'utilisation d'une *GWAS*

II. 1. *Amélioration de la densité en marqueurs grâce à l'imputation*

L'étude du phénotype a montré que ce dernier variait entre les individus et qu'il était, de plus, héritable, il est donc soumis à un déterminisme génétique. Nous avons donc cherché à savoir si des gènes pouvaient expliquer ce phénotype et pour cela, nous avons réalisé des *GWAS*.

Nous avons tout d'abord recherché une potentielle association statistique entre nos différents index et les 46 813 marqueurs (**voir Figure 33**). L'effet de chaque *SNP* a été testé en utilisant un modèle mixte univarié, implémenté dans le logiciel *GEMMA* (Genome-wide Efficient Mixed Model Association), créé par Zhou et Stephens (2012). Cette méthode a été plus amplement présentée dans la partie « Chapitre I., IV. 1. d. ». Au sein de ce modèle, une matrice centrée de parentés génomiques est utilisée et les p-valeurs calculées correspondent à des tests de Wald.

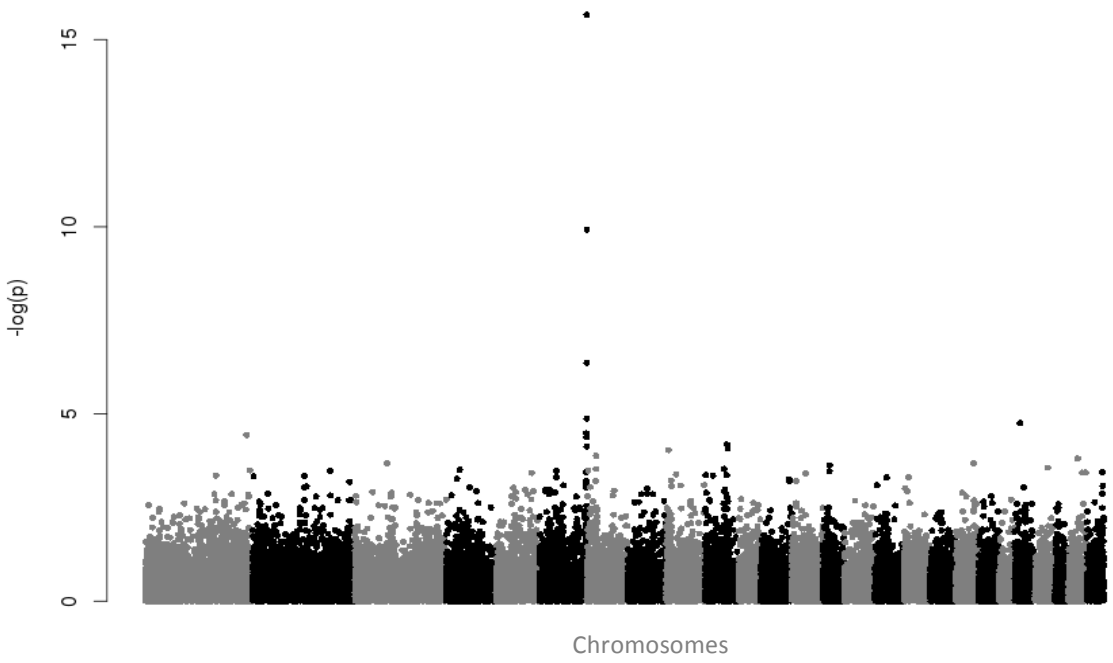

Figure 33 : Identification d'un pic avec une GWAS sur 50 000 marqueurs

Représentation des –log10(p-valeurs) pour une association simple marqueur sur les marqueurs de la puce 50K. Chaque changement de couleur correspond à un chromosome. Un pic très important est observé sur le chromosome 6.

Sur la **Figure 33**, nous remarquons que la *GWAS* sur le génome entier avec les marqueurs de la puce 50K révèle un pic assez important sur le chromosome 6 avec une p-valeur inférieure à 10^{-15}.

Nous avons voulu préciser ce résultat et pour cela, nous avons utilisé l'imputation. Les 70 Lacaune non apparentés du jeu de données populationnel ont été utilisés en tant que panel, puisqu'ils sont génotypés pour la puce haute densité 600K. Trois-cent-vingt-six pères Lacaune apparentés et génotypés pour la puce 50K ont constitué la cohorte, et donc les animaux à imputer (345 animaux moins les 19 en commun avec les 70 Lacaune non apparentés). Nous avons utilisé le logiciel BIMBAM (Guan et Stephens, 2008) pour imputer les 326 pères Lacaune sur 507 784 *SNPs*. Ce logiciel utilise le modèle de FastPhase (Scheet et Stephens, 2006) qui s'appuie sur des méthodes utilisant des groupes d'haplotypes afin d'estimer les génotypes manquants et de reconstruire les haplotypes à partir de marqueurs non phasés d'animaux non apparentés. BIMBAM a été utilisé avec les paramètres suivants :

- 10 « Expectation-Maximization (*EM*) starts ».
- Chaque *EM* est relancé 20 fois sur le panel uniquement et 1 fois supplémentaire sur la cohorte.
- Utilisation de 15 groupes.

Après l'imputation, BIMBAM estime pour chaque *SNP* de chaque animal un nombre moyen d'allèles, appelé « mean genotype ». Ce dernier est calculé à partir de la distribution *a posteriori* des trois génotypes possibles. Ces « mean genotypes » ont été montrés comme étant efficaces pour réaliser des tests d'associations (Guan et Stephens, 2008). Dans les analyses suivantes, nous utiliserons donc ces « mean genotypes » estimés par BIMBAM pour chaque marqueur de la puce 600K.

Pour valider la qualité de l'imputation, nous utilisons 10 marqueurs de la puce 600K ayant des effets significatifs après imputation : 1 issu du chromosome 6 et 9 issus du chromosome 7. Deux-cent-soixante-six animaux ont été génotypés pour ces 10 marqueurs (les seuls animaux parmi les 345 qui étaient disponibles pour cette étape de génotypage). Nous avons ainsi évalué la qualité de l'imputation en comparant les génotypes estimés aux génotypes vrais obtenus après génotypage. Pour cela, nous avons créé 10 classes (« Classe des probabilités ») sur lesquelles nous avons sommé, pour chacun des *SNPs*, les génotypes estimés par BIMBAM et les génotypes vrais donnés à l'issue du nouveau génotypage. Puis, pour chacune des classes, nous avons calculé la proportion de génotypes vrais sur les génotypes estimés (« Proportion des génotypes vrais »). Finalement, nous avons représenté cette proportion en fonction des classes choisies et nous observons une très bonne corrélation de 99% entre les deux mesures (**voir Figure 34**), ce qui valide notre imputation.

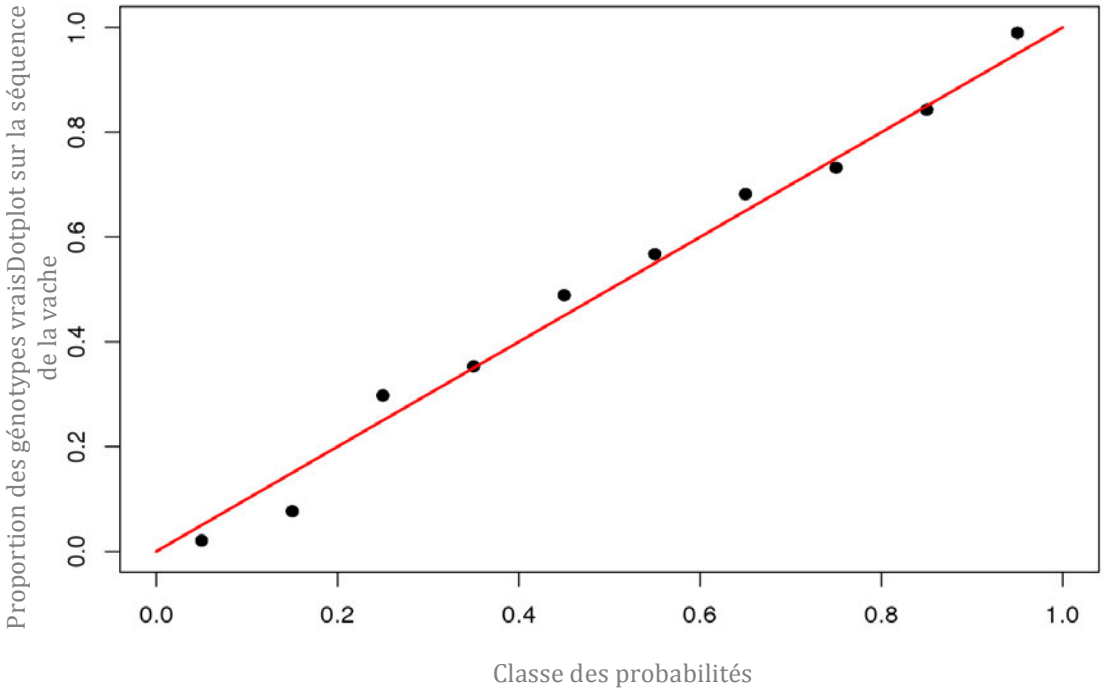

Figure 34 : Validation de l'imputation

Représentation de la proportion de génotypes vrais en fonction de leurs probabilités *a posteriori* calculées par BIMBAM.

II. 2. *La GWAS a permis d'identifier 3 QTLs*

La réalisation d'une *GWAS* sur les génotypes imputés issus de BIMBAM a permis de révéler 6 pics significatifs supplémentaires pour un *FDR* de 5% (**voir Figure 35**) : deux sur le chromosome 1, 1 sur le chromosome 6 (au même emplacement que le pic détecté précédemment), 1 sur le chromosome 7, 1 sur le chromosome 11 et le dernier sur le chromosome 19.

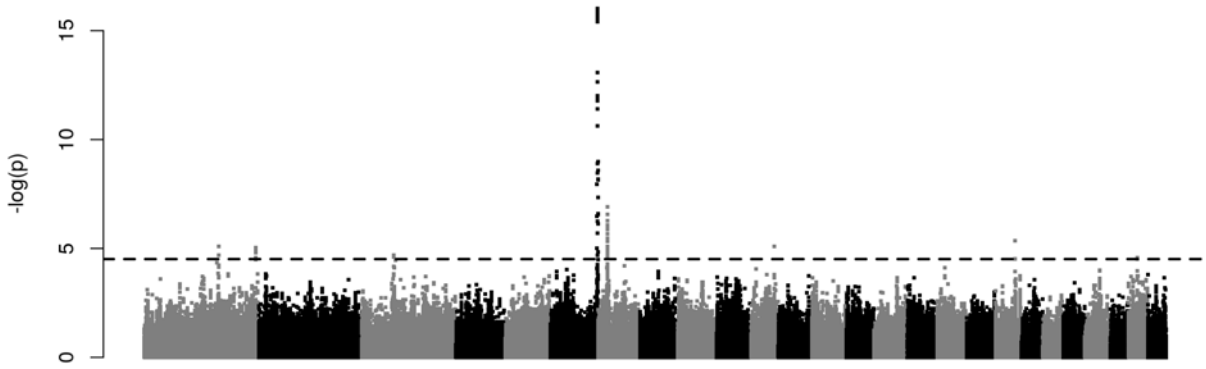

Figure 35 : Identification de 2 QTLs majeurs grâce à une GWAS

Représentation des log10 des p-valeurs pour un test d'association simple marqueur. Le niveau de significativité tout génome est donné pour une *FDR* de 5% et représenté par la droite horizontale en pointillé. L'axe des abscisses donne les chromosomes : un changement de couleur indiquant un changement de chromosome.

Les régions des chromosomes 6 et 7 présentent des p-valeurs beaucoup plus petites que celles des autres régions.

II. 2. a. *Validation de 2 QTLs grâce à l'analyse multi-QTLs*

Afin de confirmer ou d'infirmer les résultats de la *GWAS*, nous avons utilisé un modèle linéaire mixte Bayésien, qui a aussi été implémenté dans GEMMA (Zhou *et al.*, 2013). Les détails sur cette méthode se trouvent dans la partie « Chapitre 1. IV.1. d. b. ».

Suite à l'application de ce modèle Bayésien, seulement 2 régions demeurent significatives et ont donc une forte probabilité d'être des *QTLs*, tandis que le dernier pic du chromosome 1 reste suggestif et que deux régions additionnelles sont détectées sur le chromosome 3 (**voir Figure 36 et Tableau 7**). L'utilisation du modèle multi-*QTLs* permet d'estimer, qu'ensemble, les *QTLs* expliquent environ 40% de la variance génétique additive du phénotype de taux de recombinaison individuel, avec un intervalle de crédibilité de 95% allant de 28 à 53% de variance expliquée par les *QTLs*.

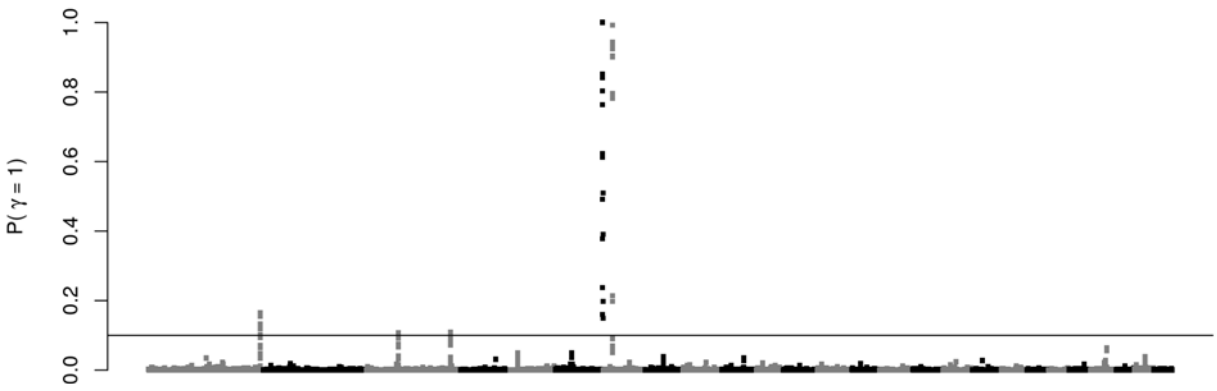

Figure 36 : Confirmation de 2 QTLs majeurs

Représentation des probabilités postérieures que des régions d'environ 20 *SNPs* soient des *QTLs* grâce à l'utilisation d'un modèle Bayésien multi-*QTLs*.

Afin de rechercher de potentiels gènes candidats dans ces régions, nous avons défini des intervalles. Pour cela, nous avons sélectionné tous les *SNPs* significatifs pour un *FDR* de 10% et nous avons défini les bornes de notre intervalle comme étant les bornes des régions englobant les *SNPs* sélectionnés. Ensuite, à l'aide de la base de données Ensembl v87 (http://www.ensembl.org/index.html), nous avons identifié les gènes présents dans cet intervalle et nous avons conservé ceux qui avaient un nom associé et qui avaient des ontologies connues.

Tableau 7 : SNPs statistiquement associés avec le phénotype du taux de recombinaison individuel

Nom du *SNP* (n° rs)	Chromo-some	Position (pb)	Allèle mineur	Fréquence	Effet β	P-valeur	p*QTL*
rs430436336	1	180 044 043	A	0,11	2,19	$8,08*10^{-6}$	0,06
rs400472211	1	268 670 581	A	0,33	0,86	$9,41*10^{-6}$	0,03
rs418551122	3	75 216 491	A	0,30	0,76	$2,42*10^{-5}$	0,04
rs407545143	3	201 298 545	G	0,24	1,13	$9,36*10^{-4}$	0,07
rs411987057	6	116 517 201	C	0,22	-2,30	$1,31*10^{-16}$	0,19
rs401206888	6	116 440 663	G	0,14	-1,95	$2,04*10^{-16}$	0,16
rs412583165	6	116 525 709	G	0,27	-2,38	$9,8*10^{-17}$	0,15
rs429477322	6	116 509 403	A	0,18	-2,17	$3,94*10^{-16}$	0,11
rs161854895	6	116 491 013	G	0,22	-2,17	$2,53*10^{-16}$	0,11
rs398811467	6	116 472 870	A	0,13	-1,94	$2,51*10^{-16}$	0,14
rs407110999	7	22 859 168	G	0,25	1,37	$8,71*10^{-7}$	0,10
rs413147562	7	22 798 236	A	0,23	1,61	$1,20*10^{-7}$	0,71

Les p-valeurs sont données pour un test simple marqueur de Wald sur un modèle animal où les *SNPs* sont considérés comme des effets fixes. ***β*** correspond à l'effet du *SNP* (en nombre de crossing-overs par méiose) sur le phénotype et ***pQTL*** correspond à la probabilité pour un *SNP* d'être un *QTL* en utilisant une méthode multi-*QTLs*.

II. 2. b. *Etude du QTL du chromosome 1*

La région significative du chromosome 1 est localisée entre 268 600 et 268 700 Kb. Chez la vache, une région orthologue, située à l'extrémité du chromosome 1 a également été montrée comme ayant un lien statistique avec le taux de recombinaison (Ma *et al.*, 2015, Kadri *et al.*, 2016). Le gène *PRDM9* a été proposé comme potentiel gène candidat. Cependant, comme nous l'avons vu précédemment, chez le mouton, *PRDM9* est situé autour de 275 000 000 Kb sur le chromosome 1, soit à environ 7 000 Kb de notre signal. Il semblerait donc que, dans notre cas, *PRDM9* ne soit pas le bon gène candidat positionnel. En revanche, sous le pic, est présent un unique gène, *KCNJ15*, qui a déjà été associé avec la réparation des *DSBs* dans les cellules humaines (Slabicki *et al.*, 2010).

II. 2. c. *Etude du QTL du chromosome 3*

Les deux régions du chromosome 3 ont été analysées. La première était localisée entre 75 162 et 75 319 Kb et contenait un seul gène codant pour le récepteur à l'hormone folliculo-stimulante *FSHR*. Bien qu'il n'affecte pas la recombinaison directement, il est nécessaire à l'initiation et au maintien d'une spermatogenèse normale chez les mâles (Tapanainen *et al.*, 1997). La deuxième région du chromosome 3 était localisée entre 201 198 et 201 341 Kb, mais ne contenait aucun gène annoté.

II. 2. d. *Etude du QTL du chromosome 7*

La région détectée sur le chromosome 7 était la deuxième région la plus significative. Elle était localisée entre 22 500 000 et 23 100 000 Kb. Tous les *SNPs* ressortant comme significatifs ont été imputés, c'est pourquoi elle n'a pas été détectée avec la *GWAS* sur les marqueurs de la 50K. La région a aussi été découverte chez les moutons Soay (Johnston *et al.*, 2016). De même, il pourrait s'agir de celle qui a été mise en évidence chez la vache (l'association étant localisée sur le chromosome 10, à environ 20 000 000 Kb). Cependant, les gènes candidats proposés dans cette espèce, *REC8* et *RNF212B* (Sandor *et al.*, 2012, Kadri *et al.*, 2016) sont situés à respectivement 2 000 000 et 1 500 000 Kb de notre pic, donc trop loin pour être à l'origine de notre signal (**voir Figure 37**). De plus, les *SNPs* ovins correspondant à ceux des bovins ne sont pas du tout significatifs dans notre analyse.

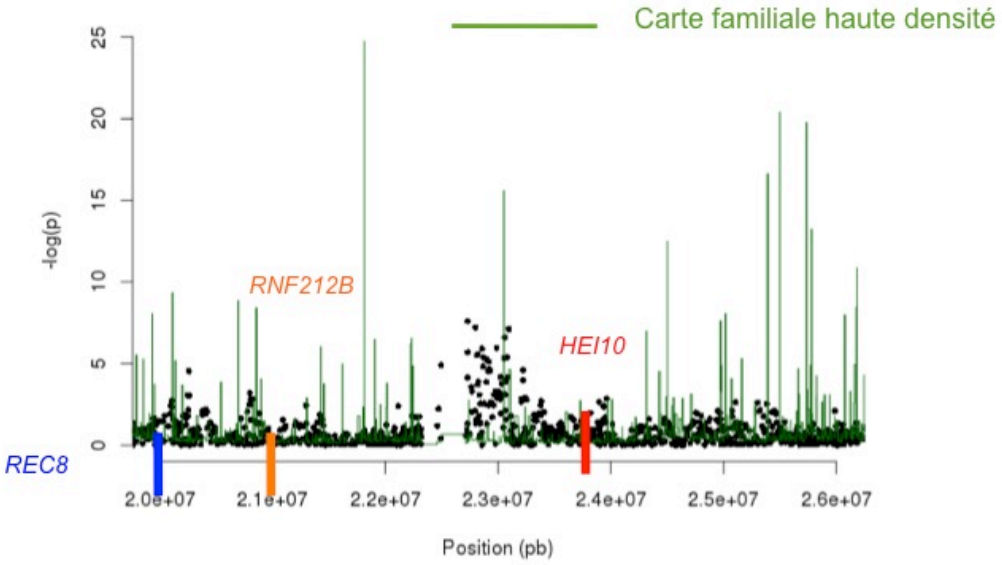

Figure 37 : Zoom sur le résultat de GWAS pour le chromosome 7

Les points noirs correspondent aux *SNPs* imputés. La carte familiale haute densité (en cM/Mb), indiquant le taux de recombinaison historique dans la zone, permet de révéler les endroits où il y a un fort déséquilibre de liaison. A proximité de notre pic, le déséquilibre de liaison est donc faible, ce qui peut être dû au *QTL* mais également au trou dans l'assemblage. Les gènes candidats *REC8* et *RNF212B* sont à près de 3 et 2 Mb respectivement de notre pic. *HEI10* semble être un candidat positionnel plus probable.

Sous notre pic, nous avons découvert 11 gènes : *OR10G2*, *OR1063*, *TRAV5*, *TRAV4*, *SALL2*, *METTL3*, *TOX4*, *RAB2B*, *CHD8*, *SUPT16H* et *RPGRIP1*. Nous avons extrait les ontologies de ces gènes à l'aide de la base de données Ensembl v87, mais aucune n'a révélé d'association avec la recombinaison. Dans la littérature un seul de ces gènes, *SUPT16H*, a été référencé comme étant impliqué dans la réparation des *DSBs* mitotiques (Kari *et al.*, 2011). En revanche, un autre gène candidat fonctionnel, *CCNB1IP1*, également appelé *HEI10*, est situé entre les positions 23 946 971 et 23 951 850 pb, soit à environ 500 Kb de notre pic. Il s'agit d'un bon candidat fonctionnel car il a été montré comme interagissant avec *RNF212* : en effet, *HEI10* permet d'éliminer les protéines *RNF212* des sites de recombinaison précoces. Il induit ainsi le recrutement d'autres intermédiaires de la recombinaison impliqués dans la maturation des crossing-overs (Qiao *et al.*, 2014, Rao *et al.*, 2016). Chez *Arabidopsis thaliana*, l'augmentation de l'expression de *HEI10* conduit à l'augmentation du nombre de crossing-overs en agissant sur les protéines *ZMM* (voir « Chapitre I. I. 4. b. a. »)., ainsi qu'à une diminution de l'interférence (Serra *et al.*, 2017).

Là-encore, les *SNPs* localisés à proximité de *HEI10* ne montrent pas d'associations significatives

avec le phénotype du taux de recombinaison individuel. Il est donc difficile de savoir clairement quel gène est le candidat fonctionnel parmi tous ces candidats fonctionnels. Cependant, nous ne pouvions pas tous les éliminer complètement. Tout d'abord, parce qu'avec seulement 345 individus, notre étude n'est peut-être pas suffisamment puissante pour localiser précisément les *QTLs*, ensuite, parce qu'il est possible que les variants causaux soient localisés à quelques centaines de kilobases du pic. Et pour finir, parce que lorsque nous étudions la région de *HEI10*, nous observons des réarrangements avec le génome humain, qui peuvent potentiellement être dus à des problèmes d'assemblage dans le génome du mouton (**voir Figure 38**).

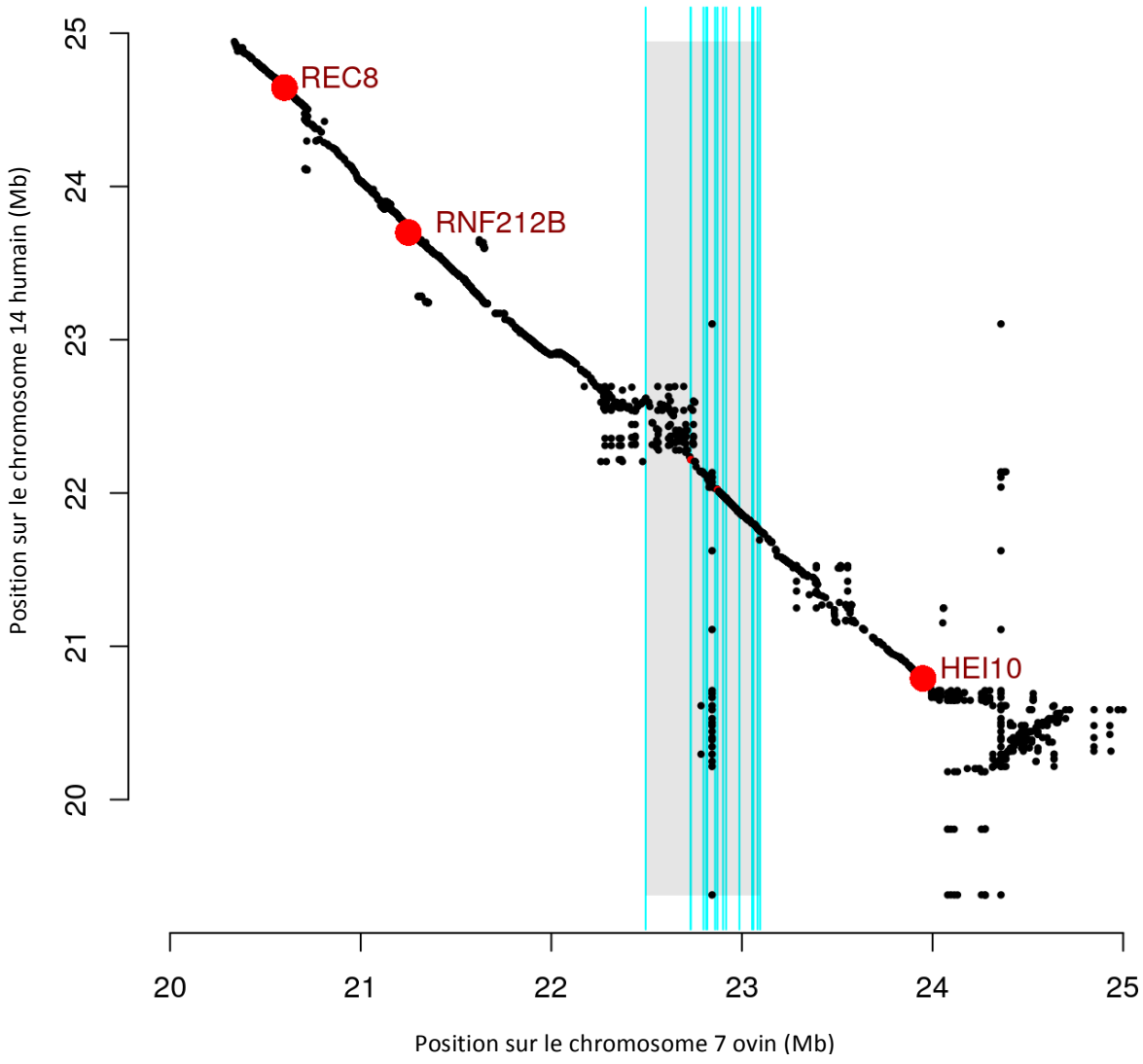

Figure 38 : Alignement local des génomes ovins et humains à proximité du QTL du chromosome 7

Réalisation d'un « dotplot » de l'alignement du chromosome 7 ovin sur le chromosome 14 humain. L'encadré gris correspond à notre pic et les barres verticales bleues correspondent aux *SNPs* significatifs. Trois gènes candidats fonctionnels associés à notre signal sont indiqués en rouge.

Ces problèmes d'assemblage pourraient être liés à la présence de séquences génomiques codant pour les chaînes alpha des récepteurs aux cellules T. Cette région est effectivement riche en séquences répétées, ce qui rend difficile son assemblage. De plus, il existe un trou, tout proche de notre pic et visible sur la **Figure 37**.

En résumé, l'identification d'un seul gène candidat fonctionnel et positionnel dans cette zone riche

en gènes et à l'assemblage douteux, n'était pas possible à partir de nos seules données.

II. 2. e. *Etude du QTL du chromosome 6*

La région la plus significative mise en évidence par la *GWAS* était localisée sur le chromosome 6 (**voir Figure 39**).

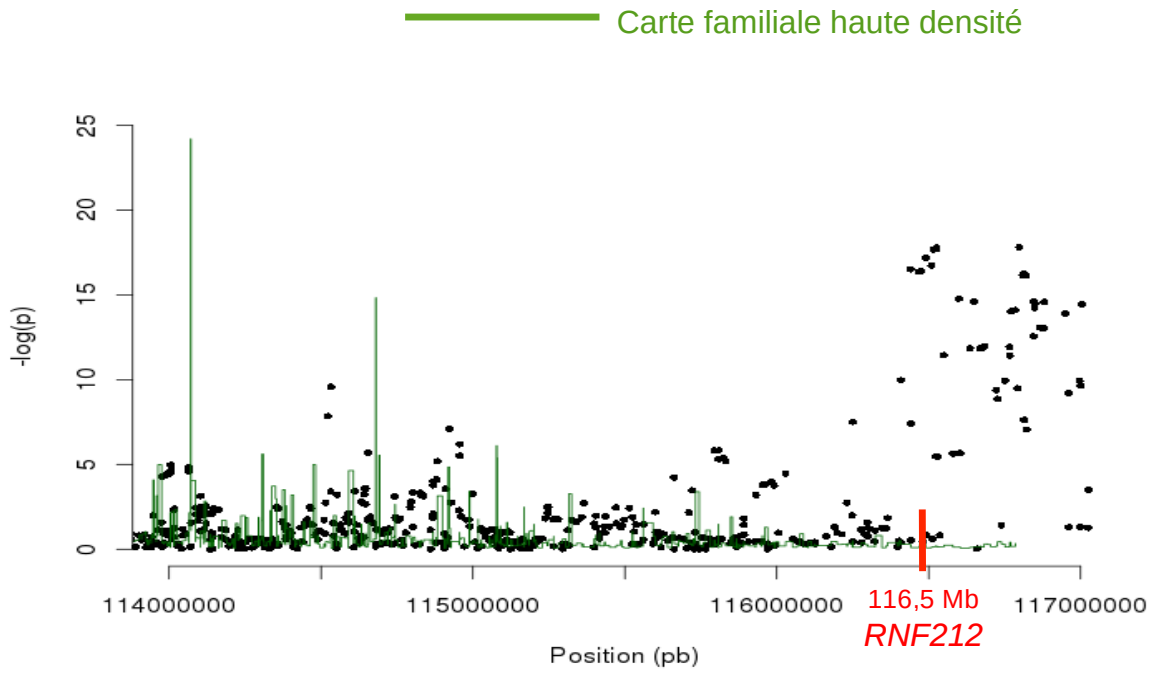

Figure 39 : Zoom sur le résultat de GWAS pour le chromosome 6

Les points noirs correspondent aux *SNPs* imputés. La carte familiale haute densité (en cM/Mb), indiquant le taux de recombinaison historique dans la zone, permet de révéler les endroits que le déséquilibre de liaison est très faible au niveau de notre pic. Le gène candidat fonctionnel *RNF212* est également un très bon candidat positionnel.

Elle correspondait à un locus déjà bien connu dans d'autres espèces et contenait 10 gènes : *CTBP1*, *IDUA*, *DGKQ*, *GAK*, *CPLX1*, *UVSSA*, *MFSD7*, *PDE6B*, *PIGG* et *RNF212*. Pour chacun de ces gènes, excepté pour *RNF212* qui n'était pas annoté sur le génome, nous avons extrait leur ontologie à partir de la base Ensembl v87, mais aucune n'était liée à la recombinaison. Cependant, deux gènes avaient déjà été reportés dans la littérature comme ayant un effet statistique sur la recombinaison : *CPLX1* et *GAK*. Kong *et al.* (2014) n'ont pas trouvé de fonction associée à la

recombinaison pour *CPLX1*, en revanche une étude a montré que la protéine *GAK* forme un complexe avec la cycline-G, complexe impliqué dans la recombinaison chez la drosophile (Nagel *et al.*, 2012). Néanmoins, au vu des études existantes, il semblerait que *RNF212* soit le meilleur candidat. En effet, chez l'Homme des variants ont été associés avec le taux de recombinaison (Kong *et al*., 2008). *RNF212* a également été associé avec la variation du taux de recombinaison chez la vache (Sandor *et al.*, 2012, Ma *et al.*, 2015, Kadri *et al.*, 2016) et chez la souris (Reynolds *et al.*, 2013). Bien que *RNF212* ne soit pas annoté sur le génome du mouton, notre pic correspond à la région bovine orthologue contenant *RNF212* (**voir Figure 40**).

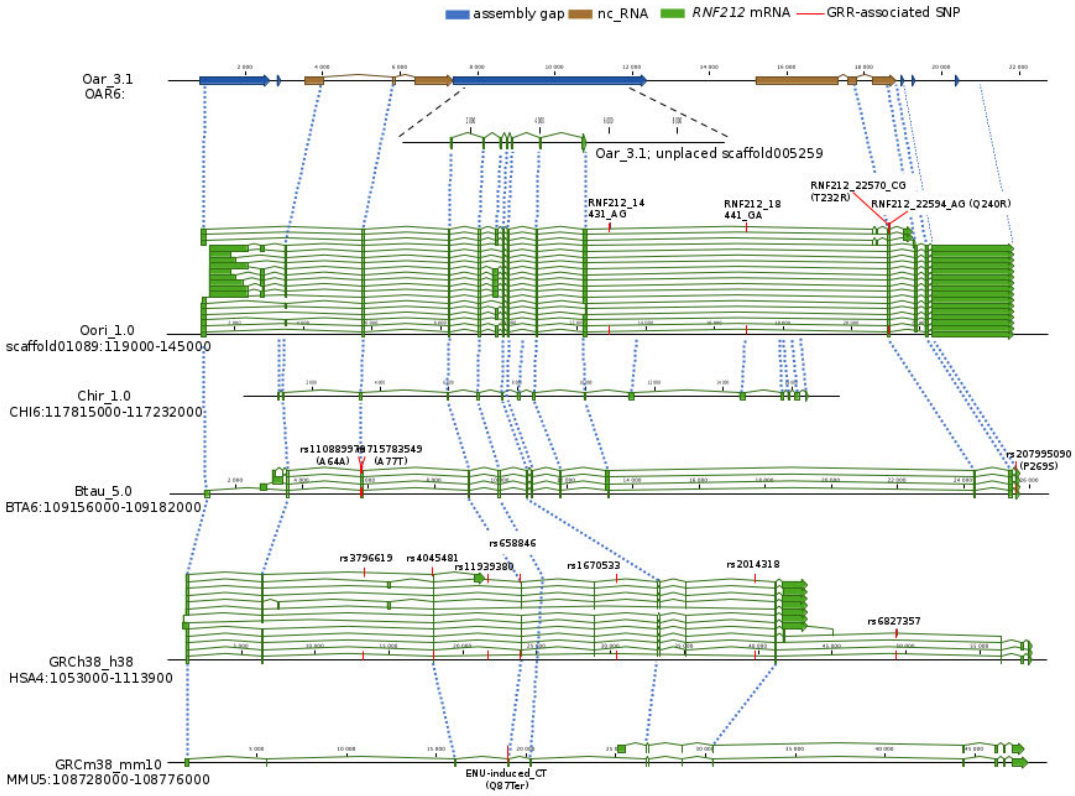

Figure 40 : Structure du gène RNF212 dans des espèces variées

Plusieurs séquences non codantes d'*ARN* (nc_RNA en marron) font partie de la séquence du gène *RNF212*. Le gène ovin est également présent sur le « scaffold00529 » non assemblé, mais qui peut être virtuellement localisé dans le trou de l'assemblage (en bleu). Chez *Ovis orientalis*, le gène *RNF212* comporte 14 exons avec des épissages alternatifs (*ARNm* en vert). Les analyses d'homologie (traits en pointillés) avec les gènes *RNF212* annotés chez les autres ruminants (chromosomes 6 de la vache et de la chèvre) montrent une bonne conservation de la structure du gène avec la vache, mais une conservation plutôt partielle avec la chèvre, où seulement les 6 derniers exons correspondent aux régions introniques du mouton. Lorsque nous comparons avec le gène *RNF212* humain, situé sur le chromosome 4, et celui de la souris, situé sur le chromosome 5, seulement 5 exons sont conservés avec les ruminants, ce qui montre que la structure du gène ne semble pas être conservée. Les lignes rouges localisent les *SNPs* associés avec le taux de recombinaison découverts dans cette étude, et dans les études précédentes.

II. 3. *Recherche de mutations candidates dans le gène RNF212*

Afin de vérifier que *RNF212* est un gène candidat positionnel pertinent, nous avons étudié l'association de ses polymorphismes avec le phénotype du taux de recombinaison individuel.

II. 3. a. *Identification et assignation de la séquence ovine de RNF212*

Le gène n'est pas encore annoté sur le génome de référence du mouton ; *Ovis aries* v3.1. Néanmoins, une séquence complète du gène a pu être mise en évidence dans un « scaffold » non assigné de l'espèce ovine *Ovis orientalis musimon* ; « scaffold01089 ». Grâce à l'utilisation du logiciel Blast, nous avons réalisé des alignements qui ont montré que ce scaffold pouvait être placé précisément au niveau de notre zone *QTL*, entre les positions 116 426 000 et 116 448 000 pb chez le mouton *Ovis aries*. La position a été confirmée à l'aide d'alignements avec le génome bovin. Le gène *RNF212* ovin mesure donc 23,7 Kb et est composé de 12 exons. Cependant, des *ARN* messagers du gène révèlent de multiples exons alternatifs. Un autre « scaffold », présent sur le génome d'*Ovis aries* cette fois-ci ; « scaffold005259 », contenait la partie centrale de *RNF212* (les exons 4 à 9) et pouvait être placé au niveau d'un large trou dans l'assemblage du génome ovin. Pour finir, des *ARNs* non codants de *RNF212* correspondaient parfaitement à des séquences exoniques de *RNF212* (**voir Figure 40**). Il est intéressant de noter que la structure ovine de *RNF212* ne semble pas bien conservée avec celle de la chèvre, de l'Homme ou de la souris.

II. 3. b. *Découverte de polymorphismes de RNF212 dans la population Lacaune*

A partir de la séquence génomique et de la structure de *RNF212* découvertes chez *Ovis aries orientalis*, nous avons créé de nombreuses amorces grâce au logiciel PRIMER3. Elles ont permis d'amplifier chaque exon annoté chez le mouton et quelques introns ovins correspondant à des régions exoniques chez la chèvre *Capra hircus*. Les amplifications ont été réalisées par *PCR*. Chaque produit d'amplification à été généré à partir de 50 ng d'*ADN* provenant de 4 individus Lacaune homozygotes avec des génotypes extrêmes pour le meilleur *SNP* de la zone *QTL* du chromosome 6 obtenu sur la puce 50K (donc non imputé) ; *SNP* rs418933055, p-valeur de $2,56*10^{-17}$. Suite à cette étape d'amplification, les produits de *PCR* ont été séquencés, analysés, puis alignés contre la séquence de *RNF212* d'*Ovis orientalis* afin d'identifier d'éventuels

polymorphismes (**voir Annexe 1 pour les détails de la réalisation**). Quatre mutations ont ainsi été mises en évidence : deux *SNPs* dans l'intron 9 et deux *SNPs* dans l'exon 10 ; une mutation synonyme et l'autre qui modifiait effectivement un acide aminé.

II. 3. c. *Génotypage des mutations de RNF212*

Les 266 animaux pour lesquels nous avions l'*ADN* disponible ont été génotypés pour ces 4 mutations à l'aide de *RFLP*, après une étape d'amplification par *PCR* et de digestion par des enzymes de restriction (**voir Annexe 1 pour les détails de la réalisation**).

Nous avons ensuite testé l'association de ces mutations avec notre phénotype grâce à une *GWAS* et nous avons calculé le déséquilibre de liaison entre ces mutations et le *SNP* le plus associé sur la puce 600K (**voir Tableau 8**).

Tableau 8 : Mutations détectées dans le gène RNF212

Nom de la mutation	Changement de base	Position sur le génome v3.1 du mouton	Position sur le génome d'*Ovis orientalis*	Reconstruction des positions sur le génome v3.1	Fréquence	Effet β	P-valeur	p*QTL*
RNF212_14431_AG	A > G	Contig Un_JH922970 : 5925	Scaffold01089 : 132 229	6 : 116 786 130	0,18	-3,98	$6,25*10^{-17}$	0,23
RNF212_18411_GA	G > A	6 : 116 438 363	Scaffold01089 : 136 209	6 : 116 825 930	0,17	-5,58	$4,93*10^{-15}$	0,02
RNF212_22570_CG	C > G	6 : 116 442 942	Scaffold01080 : 140 368	6 : 116 867 520	0,18	-3,94	$4,61*10^{-16}$	0,09
RNF212_22594_AG	A> G	6 : 116442 966	Scaffold01089 : 140 392	6 : 116 867 760	0,17	0,57	0,54	0,004

Polymorphismes détectés dans le gène *RNF212* après le séquençage d'animaux clés. Les positions sur le génome v3.1 du mouton correspondent aux positions réelles des mutations sur le génome du mouton, avant intégration du « scaffold » d'*Ovis orientalis* ; la première mutation est la seule qui se situe dans un « contig » non assemblé. Les positions reconstruites correspondent aux positions supposées des mutations sur le génome du mouton après intégration du « scaffold ». Les p-valeurs sont données pour un test de Wald, ***β*** correspond à l'effet de la mutation sur le phénotype et ***pQTL*** correspond à la probabilité pour la mutation d'être un *QTL*.

Deux de ces mutations étaient très associées avec le taux de recombinaison individuel, leur p-valeur étant du même ordre de grandeur ($p < 10^{-16}$) que le *SNP* imputé le plus fort (rs412583165),

l'une d'entre elles était même plus significative que le *SNP* imputé le plus significatif (p-valeur de $6,25*10^{-17}$ vs p-valeur de $9,8*10^{-17}$). Cependant, cette mutation était intronique et ne semblait pas avoir d'effet sur la protéine. Nous observons une très forte adéquation entre les valeurs de déséquilibre de liaison entre une mutation et le meilleur *SNP* et leur p-valeur associée (**voir Figure 41**).

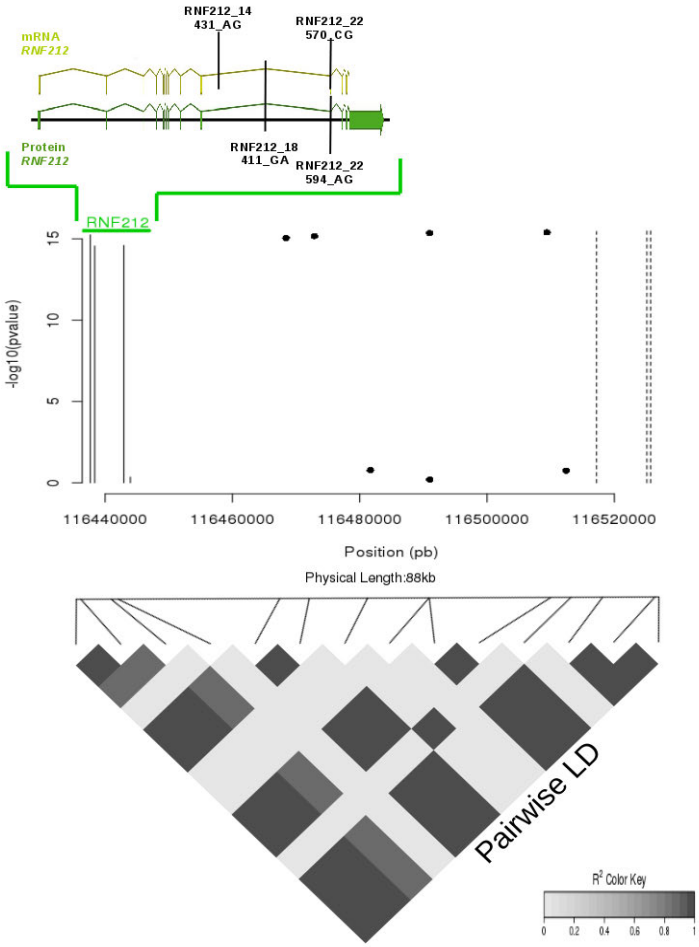

Figure 41 : Déséquilibre de liaison entre les polymorphismes du gène RNF212 et les SNPs du QTL du chromosome 6

La figure du haut représente les *ARNm* et la protéine *RNF212*. Les 4 mutations génotypées sont indiquées : les deux premières sont introniques et les deux autres sont exoniques. On replace le gène sur un zoom du *QTL* du chromosome 6 (figure du milieu). Les 4 lignes verticales pleines représentent les mutations, tandis que les lignes verticales pointillées représentent les 3 *SNPs* les plus significatifs. Les points du milieu montrent les *SNPs* intermédiaires entre les mutations et les *SNPs* significatifs. Finalement, la figure du bas, indique le déséquilibre de liaison entre les mutations et tous les *SNPs* présentés sur la figure du milieu. Cela révèle deux blocs d'haplotypes : un entre les 3 mutations les plus significatives et un autre entre les 3 *SNPs* les plus significatifs.

En résumé, ces résultats montrent que des polymorphismes de *RNF212* sont très fortement associés au phénotype du taux de recombinaison individuel. Ceci confirme donc que *RNF212* n'est pas seulement un gène candidat fonctionnel, c'est aussi un très fort candidat positionnel.

III. Le déterminisme génétique du taux de recombinaison individuel diffère entre les Soay et les Lacaune mâles

Comme pour l'étude des cartes de recombinaison, nous avons souhaité comparer nos résultats de déterminisme génétique avec ceux obtenus chez les Soay (Johnston *et al.*, 2016).

L'utilisation de *GWAS* chez le Soay a permis d'identifier deux *QTLs* majeurs, qui semblent avoir des effets sexe-spécifiques. Ces deux *QTLs* sont localisés dans les mêmes régions que celles que nous avons déterminées, sur les chromosomes 6 et 7. Cependant, le *QTL* du chromosome 6 en Soay ne semble avoir un effet que chez les femelles, or nous n'avons réalisé notre *GWAS* que sur des Lacaune mâles.

De plus, bien que les *QTLs* soient localisés dans des régions similaires, les *SNPs* les plus significatifs sont différents entre les deux espèces (**voir Figure 42**).

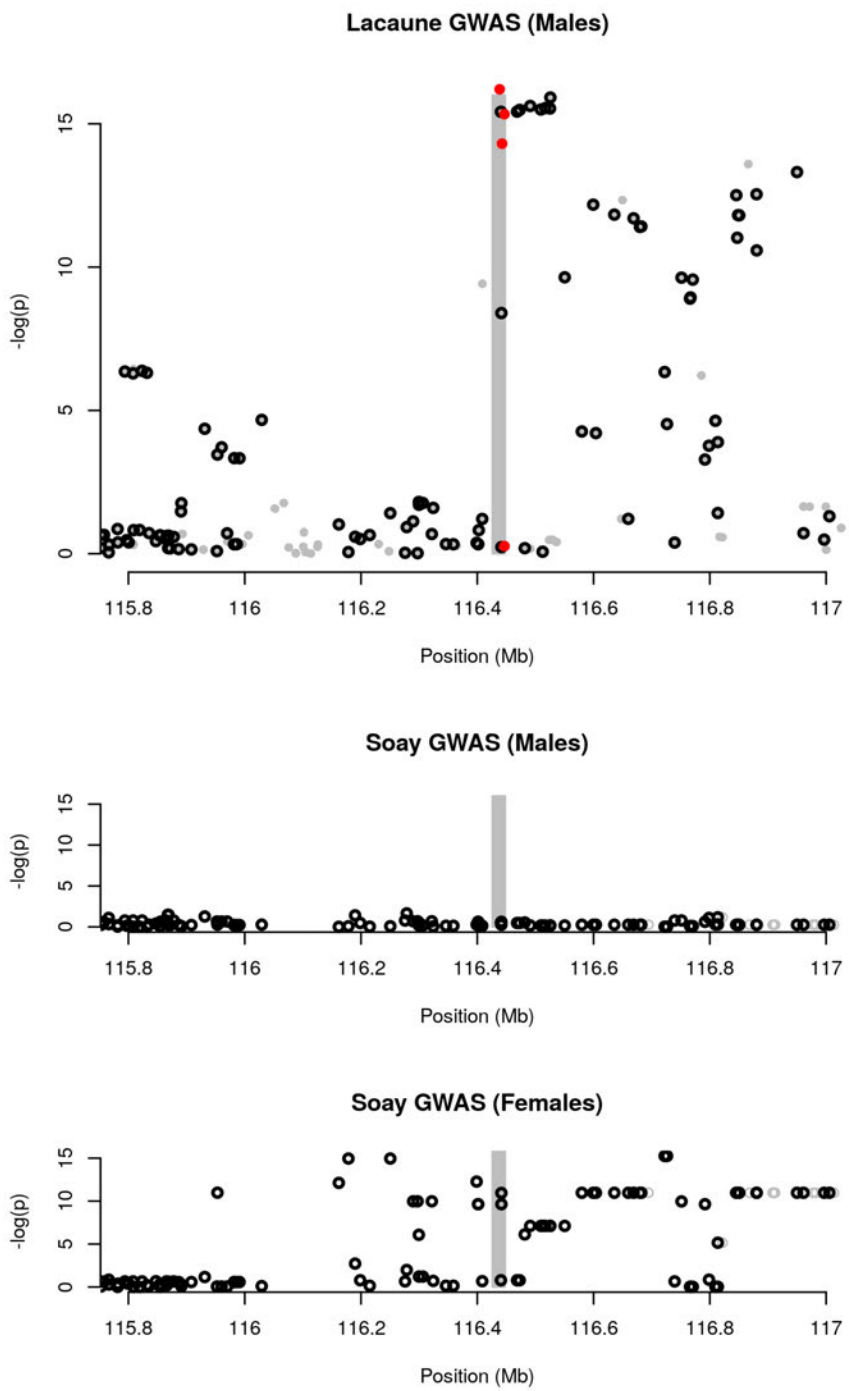

Figure 42 : Comparaison des résultats de GWAS pour le chromosome 6 entre les Lacaune mâles (graphe du haut), les Soay mâles (graphe du milieu) et les Soay femelles (graphe du bas)

La zone sombre indique la position prédite de *RNF212*. Les ronds noirs sont les marqueurs testés dans les deux populations. Les points rouges sont les mutations au sein de *RNF212* découvertes dans notre étude et génotypées dans la population Lacaune. Les résultats des *GWAS* en Soay proviennent de l'étude de Johnston *et al.* (2016).

Ceci peut s'expliquer de deux façons ; soit les deux populations ont les mêmes *QTLs* qui ségrégent, mais les *SNPs* non similaires des *GWAS* sont dus à des patrons de déséquilibre de liaison différents entre les deux races ; soit les deux populations ont des mutations causales différentes dans la même région. Cependant, étant donné que les deux races sont toutes deux des moutons domestiques, qui ne se sont pas séparés depuis très longtemps, nous avons tendance à favoriser la première hypothèse. Densifier les données de génotypage, par exemple en génotypant les mutations de *RNF212* découvertes dans la race Soay, permettrait d'avoir une réponse plus claire.

Nous avons également remarqué que, dans notre étude et dans celle des Soay, le *QTL* du chromosome 7 n'a pas été découvert avec une simple *GWAS* sur la puce 50K. Il a fallu que nous utilisions l'imputation pour le révéler et Johnston *et al*. (2016) ont utilisé une méthode de cartographie par héritabilité régionale, une approche en composantes de variance qui estime la variance additive expliquée par des régions génomiques d'une taille donnée (Nagamine *et al.*, 2012). Une telle observation rend plus difficile de savoir s'il s'agit d'une même mutation causale ou de différentes pour une même localisation dans les deux populations.

Pour ce qui est des gènes candidats, en Soay, ce sont les gènes *CPLX1* et *GAK* qui ont été proposés comme candidats pour le *QTL* du chromosome 6, sûrement car *RNF212* n'est pas annoté sur le génome, or il ne se situe qu'à 177 Kb de leur locus. Et pour le chromosome 7, ce sont les mêmes que ceux proposés chez la vache : *RNF212B* et *REC8*.

IV. Conclusions et perspectives

Au cours de cette thèse, nous avons cherché à savoir si le taux de recombinaison individuel était également soumis à un déterminisme génétique chez le mouton. Pour cela, nous avons utilisé le jeu de données familiales afin de déterminer une valeur de phénotype pour chacun des 345 pères, puis nous avons réalisé des *GWAS* sur ces phénotypes, ce qui a permis de mettre en évidence 2 gènes candidats majeurs.

IV. 1. *Estimation du taux de recombinaison comme un index*

Afin d'étudier le déterminisme génétique de la variation du taux de recombinaison en Lacaune, nous avons tout d'abord estimé l'héritabilité de ce phénotype, en utilisant une approche

classique sur un grand pédigrée. Grâce à cela, nous avons pu extraire les valeurs génétiques additives du phénotype pour chacun des 345 pères. Ces valeurs ont ensuite été utilisées pour réaliser des *GWAS*. Les valeurs génétiques additives sont, par définition, seulement estimées à partir de facteurs génétiques, étant donné que les effets environnementaux ayant un impact sur le phénotype ont été supprimés. Et, en effet, nous avons remarqué que la proportion de variance de ces valeurs génétiques additives expliquée par les facteurs génétiques était proche de 1. Ainsi, malgré la taille de notre échantillon qui peut limiter la résolution des analyses, la puissance de notre *GWAS* est grandement augmentée grâce à la très bonne précision du phénotype.

L'héritabilité du taux de recombinaison individuel en Lacaune a été estimée à 0,23, ce qui est similaire aux valeurs obtenues dans d'autres études. Ainsi, chez la vache, l'héritabilité était de 0,22 (Sandor *et al*., 2012), mais celle des mâles Soay est plus faible que la nôtre : 0,12 (Johnston *et al.*, 2016). Nous discuterons de cette différence plus loin. Il y a peu d'informations disponibles dans la littérature sur l'impact des facteurs environnementaux sur le taux de recombinaison, cependant, nous avons trouvé un effet significatif du mois d'insémination sur le taux de recombinaison individuel, en particulier avec le mois de Mai. Une confirmation de cet effet et une interprétation biologique de ce résultat nécessiteraient une étude dédiée. Cependant, cela semble cohérent avec le fait qu'en Lacaune, c'est de la semence fraîche qui est utilisée pour les inséminations et de plus, la reproduction de cette espèce est saisonnée (Rosa et Bryant, 2003).

IV. 2. *Comparaison des QTLs avec les autres espèces*

IV. 2. a. *La variation de la recombinaison semble être sous un déterminisme polygénique*

Le déterminisme génétique découvert dans notre étude ressemble fortement à celui des autres espèces, en particulier les ruminants. Deux *QTLs* majeurs affectent la variation du taux de recombinaison en Lacaune et sont communs à ceux de la vache et du mouton Soay. Les gènes et mutations candidates à l'origine de ce déterminisme ne sont, pour l'heure, pas encore découverts. Cependant les deux régions contiennent des gènes impliqués dans la constitution des crossing-overs : *RNF212* et *HEI10* (Qiao *et al.*, 2014, Rao *et al.*, 2016) ; ces deux gènes étant donc de sérieux

candidats fonctionnels. Un troisième gène a été identifié dans notre étude, *KCNJ15*. C'est un nouveau candidat, encore non proposé pour être à l'origine de ce phénotype. Ses rôles et mécanismes d'action dans la réparation des *DSBs* doivent encore être confirmés et étudiés. Il est donc intéressant de noter que ces trois gènes sont liés à la réparation des *DSBs* et au processus de création des crossing-overs. Pour finir, le quatrième gène candidat, *FSHR*, bien que largement documenté sur ces effets sur la gamétogenèse, ne semble apparemment pas lié directement à la recombinaison.

Dans notre étude, 60% de la variance génétique additive du taux de recombinaison individuel n'est pas expliquée par les *QTLs* et est due à des effets polygéniques. Ceci peut s'interpréter à l'aide d'études récentes qui ont montré que d'autres mécanismes, notamment impliqués dans l'évolution de la conformation des chromosomes au cours de la méiose, pouvaient expliquer une part importante de la variation du taux de recombinaison entre différentes lignées de souris (Baier *et al.*, 2014) et des espèces bovines (Ruiz-Herrera *et al.*, 2017). De plus, les variations au niveau des principaux loci liés à la recombinaison chez les Mammifères (*RNF212*, *CPLX1*, *REC8* ou encore l'inversion *17q21.31*) n'expliquent que 3 à 11% de la variance phénotypique parmi les individus (Ritz *et al.*, 2017). Un tel résultat favorise l'hypothèse que le taux de recombinaison est gouverné par des mécanismes biologiques multiples et indépendants, qui ont des déterminismes génétiques distincts.

Nos résultats indiquent effectivement que le déterminisme génétique de ce phénotype semble plutôt de nature polygénique. Afin de mieux le comprendre, il faudrait utiliser des échantillons de grande taille ou combiner différentes approches (Baier *et al.*, 2014, Ruiz-Herrera *et al.*, 2017).

IV. 2. b. *Le taux de recombinaison est contrôlé par différents mécanismes chez les Soay*

Notre étude est la seconde analyse menée sur le déterminisme génétique chez le mouton. La première ayant été réalisée chez les Soay (Johnston *et al.*, 2016).

La combinaison des deux jeux de données issus de la race Lacaune et la race Soay, permet d'étudier l'évolution de la recombinaison sur une période de temps relativement courte. L'une des

plus importantes différences entre les deux études, est que les deux *QTLs* détectés n'ont pas d'effet en Soay mâles, alors qu'ils ont de très forts effets en Lacaune mâles. Cela pourrait être dû à une moins bonne résolution en Soay. En effet nous avons près de 170 000 crossing-overs supplémentaires en Lacaune mâles (**voir Tableau 4**). Cependant, les deux populations ont des héritabilités polygéniques similaires ; sachant que les *QTLs* des Lacaune expliquent environ 40% de la variance génétique additive, il est possible d'estimer la variance génétique additive polygénique des Lacaune mâles. Pour cela, nous partons de la formule donnant l'héritabilité h^2 :

$$h^2 = \sigma_A^2 / (\sigma_A^2 + \sigma_E^2) \quad (1)$$ avec :

- σ_A^2 : variance génétique additive.
- σ_E^2 : variance environnementale.

L'héritabilité polygénique h_p^2 est donnée par la formule suivante :

$$h_p^2 = \sigma_P^2 / (\sigma_P^2 + \sigma_E^2) \quad (2)$$ avec :

- σ_P^2 : variance polygénique.

Or la variance génétique additive σ_A^2 peut être exprimée selon la variance polygénique σ_P^2 et la variance des *QTLs* σ_Q^2:

$$\sigma_A^2 = \sigma_Q^2 + \sigma_P^2 \quad (3)$$

Et la variance des *QTLs* est également liée à la part de variance expliquée par les *QTLs*, ***PGE***, selon la façon suivante :

$$PGE = \sigma_Q^2 / (\sigma_Q^2 + \sigma_P^2) \quad (4)$$

En combinant les équations **(1)**, **(2), (3) et (4)** nous obtenons σ_P^2 et de là nous en déduisons h_p^2 :

$$\sigma_P^2 = \sigma_A^2 * (1 - PGE)$$

La variance génétique additive polygénique des Lacaune mâles est ainsi estimée à 0,16, ce qui est du coup très similaire à la valeur de 0,12 déterminée chez les Soay mâles. De plus, les deux cartes génétiques sont très similaires, tant en terme d'intensité, qu'en terme de distribution sur le génome. Ceci peut suggérer que les patrons de recombinaison sont conservés entre les races, mais ont des déterminismes génétiques distincts. Des travaux supplémentaires seront nécessaires afin d'avoir une meilleure compréhension du contrôle génétique de la recombinaison chez le mouton ; ils pourraient notamment associer des techniques génétiques, cytogénétique, moléculaires et bio-

informatiques.

Bien que nous ayons découvert les deux mêmes principaux *QTLs* sur les chromosomes 6 et 7, d'autres *QTLs* ont également été découverts dans les deux races. Ainsi, Johnston *et al*. (2016) ont également mis en évidence un *QTL* sur le chromosome 3, tant chez les mâles que chez les femelles, pour lequel ils n'avaient cependant pas de gènes candidats fonctionnels. Il n'est cependant pas localisé aux mêmes endroits que les nôtres. On peut ainsi se demander pourquoi il n'a pas été retrouvé chez les Lacaune mâles. Il pourrait ainsi s'agir d'une région spécifique de la race Soay ou bien un artefact qui a conduit à un faux positif. Le même raisonnement peut également s'appliquer avec notre *QTL* du chromosome 1 et nos régions du chromosome 3 qui n'ont pas été retrouvées chez les Soay, mâles ou femelles.

IV. 3. *La recherche de mutations causales*

Etant donné que le gène *RNF212* était sans doute notre candidat fonctionnel le plus solide, nous avons décidé de rechercher de potentiels polymorphismes liés à la variation du taux de recombinaison dans ce gène. Nous avons ainsi découvert 4 mutations, dont 3 très fortement associées au meilleur *SNP* et au phénotype. Cependant, au vu des p-valeurs obtenues, il semblerait que nous n'ayons pas trouvé la mutation causale.

Il pourrait également être intéressant de faire cette recherche pour les *QTLs* des chromosomes 1 et 7. Cependant, c'est un travail très lourd. En effet, il faut pouvoir séquencer des animaux extrêmes, donc il faut avoir à disposition de l'*ADN*, créer plusieurs amorces et réaliser des *PCRs* qui présentent toutes deux des limites et des risques de faux positifs. A la suite de cela, il faut pouvoir génotyper les animaux pour les mutations, or les génotypages échouent parfois. De plus, les mutations découvertes sont très souvent introniques ; il peut donc être difficile de relier leur potentiel effet biologique au phénotype étudié.

Il s'agit donc de techniques lourdes, longues et coûteuses à mettre en œuvre, qu'il vaut donc mieux réserver pour des candidats solides.

IV. 4. *Nouvelle approche pour l'étude du déterminisme grâce à l'utilisation d'une race croisée, la Romane*

La race Romane, anciennement INRA401, est une race récente, créée par l'*INRA* dans les années 70 afin d'augmenter la productivité du troupeau ovin français. La race est ainsi issue de croisements successifs entre une race hyper-prolifique, la Romanov et la race française bouchère par excellence, le Berrichon du Cher. Etant donné que la race est issue d'un croisement récent, ses chromosomes sont des mosaïques des chromosomes ancestraux provenant des races d'origine. Il est possible de reconstituer ces mosaïques à partir de données de génotypage dense et une fois cette reconstitution effectuée, nous pouvons exploiter la distribution des jonctions entre segments de croisement afin d'estimer le taux de recombinaison à l'échelle du génome.

Dans un projet démarré lors de la fin de ma thèse, *Romane Ite Domum* (**voir Annexe 2**), et qui en sera donc une suite logique, des grandes familles de Romane génotypées sur la puce haute densité 600K vont être utilisées afin d'améliorer la précision de localisation des crossing-overs. Ce qui permettrait par la suite de calculer précisément la proportion de crossing-overs qui tombe dans les points chauds, à l'échelle de la kilobase et plus de la mégabase, comme c'est le cas actuellement avec les données Lacaune.

Ces nouvelles données permettront donc de disposer d'une collection de crossing-overs bien mieux résolus qu'en Lacaune (de l'ordre de la centaine de Kb) pour un petit nombre d'animaux (**voir Figure 43**).

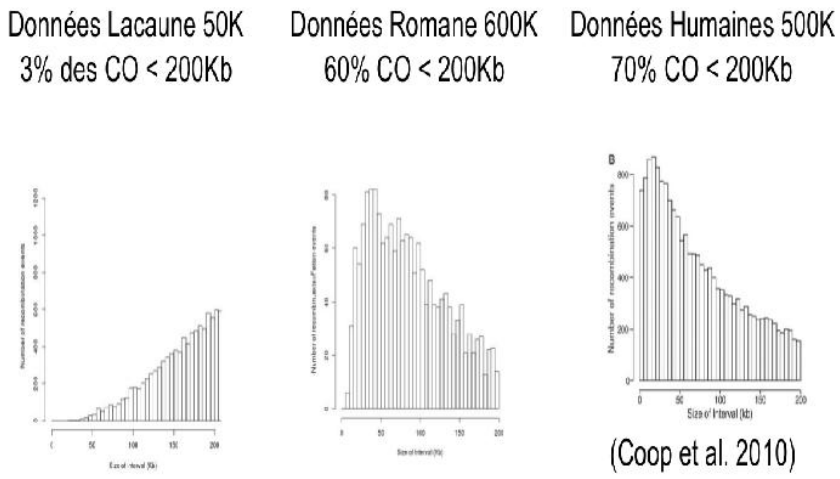

Figure 43 : Distribution de la taille des intervalles de résolution des crossing-overs détectés dans 3 dispositifs.

Un phénotypage fin des individus pour le biais d'usage des points chauds sera ainsi obtenu. Cependant, il n'y aura que quelques dizaines d'individus phénotypés, ce qui ne permettra probablement pas de recherches de *QTLs*. Pour autant, le gène *PRDM9* étant le plus connu pour être à l'origine de ce phénotype, les analyses se focaliseront sur ce gène.

Il sera également possible de collecter de nombreux haplotypes de la race Romane grâce à la reconstruction des haplotypes parentaux, ce qui permettra d'établir une carte de recombinaison basée sur les patrons de croisement (**voir Figure 44**).

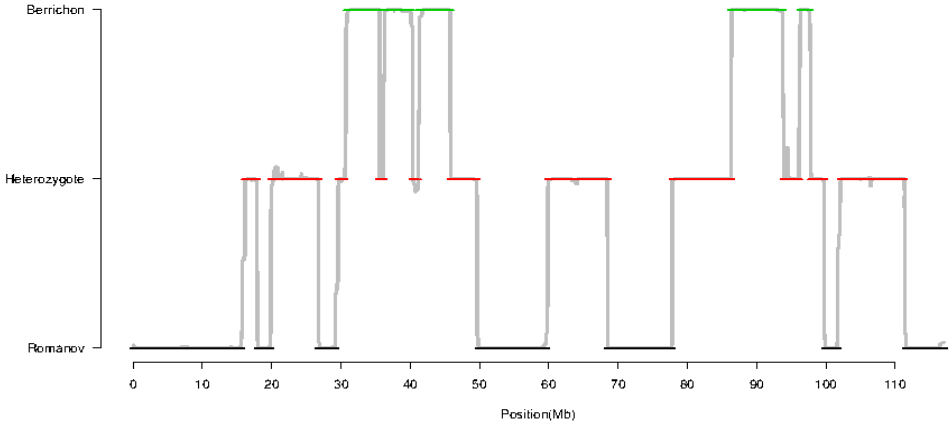

Figure 44 : Reconstruction des origines populationnelles

Exemple de reconstruction des origines populationnelles (Berrichon ou Romanov) des haplotypes d'un individu Romane génotypé avec la puce HD sur le chromosome 6 ovin.

La reconstruction des haplotypes parentaux permettra également d'enrichir la collection des haplotypes 600K disponibles dans la race d'origine ; Romanov et Berrichon du Cher, et ainsi de détecter des points chauds de recombinaison dans les deux races, par une approche populationnelle, comme celle menée en Lacaune.

Chapitre 4 : Création de Puces Basse-Densité

Chapitre 4 : Utilisation des cartes génétiques en sélection : création de puces basse densité

Cette partie fera l'objet d'une prochaine publication et est donc en cours de rédaction, en particulier la partie « Discussion ».

L'étude de la recombinaison en Lacaune a mis à disposition des outils, notamment les cartes génétiques. Nous avons donc réfléchi à une utilisation concrète de ces cartes génétiques en sélection génomique (cf « Chapitre 2. I. 1. b » pour les détails de la mise en place de la sélection génomique en Lacaune).

Depuis quelques années, la sélection génomique cherche à permettre le génotypage d'un grand nombre d'individus mais pour un faible coût. Pour cela, le principe de l'imputation a été développé. Il s'agit d'utiliser des puces basse densité pour pouvoir génotyper un plus grand nombre d'animaux. La qualité des analyses génomiques reste toutefois conservée car il est possible d'imputer ces animaux sur des puces moyenne ou haute densité. En Lacaune, une telle puce, d'environ 16 000 marqueurs, a été développée en 2015 et l'imputation des animaux génotypés sur cette puce sur la puce moyenne densité 50K n'affecte pas la qualité de l'indexation génomique (Larroque *et al.* In Prep).

Actuellement, la plupart des puces basse densité sont créées à partir des distances physiques entre marqueurs, afin qu'ils soient régulièrement espacés sur le génome. La mise à disposition des cartes génétiques en Lacaune nous a ainsi permis d'étudier l'impact sur la qualité d'imputation de leur utilisation pour créer ces puces basse densité.

Le contexte, les matériels et méthodes utilisés, ainsi que les résultats et la dicussion sont présentés en détails dans l'article en préparation suivant.

Comparison of imputation accuracy using *SNPs* sets based on a physical map or on a genetic map

Morgane Petit*, Jean-Michel Astruc†, Marjorie Chassier*, Hélène Larroque*, Bertrand Servin* and Carole Moreno*

* INRA, Génétique, Physiologie et Système d'Elevage, F-31326 Castanet-Tolosan, France

† Institut de l'Elevage, F-31326 Castanet-Tolosan, France

Abstract

The imputation is a robust tool able to infer the genotypes at un-genotyped loci based on known genotypes in the vicinity. It allowed to predict genotypes of thousands of *SNPs* with a low-density array, which is less expensive and thus can improve the genomic selection, while maintaining good genotyping information thanks to the imputation of this low-density array on a medium or a high density array. In Sheep, 4 *SNPs* arrays do exist: a 300 *SNPs* array, a low-density array (16K *SNPs*), a medium density array (54K *SNPs*) and a high-density array (600K *SNPs*). The criteria to select the *SNPs* to put on the arrays were similar: regular physical distances between *SNPs* and the highest *SNP* Minor Allele Frequency (*MAF*). Usually, in genomic selection, young candidates rams were genotyped with a low-density array and a genomic prediction is performed after imputation of the unknown *SNPs* of a medium density array using *SNPs* of the low-density array. Consequently, a first selection of rams can be performed before the entrance in the control station. In this paper, the aim is to test by simulations the imputation accuracy with different *SNPs* density (300, 3K, 10K, 16K) when *SNPs* are selected based on the genetic distances or the physical

distances. For a selection of 10K *SNPs*, the use of genetic distances improves the imputation accuracy of above 20% in average, especially at the extremities of the chromosomes, where there are more and better located markers selected from the genetic map, than from the physical map. The selection of 3K gives intermediate results, with a better imputation accuracy at the extremities of the chromosomes. For a selection of 300 *SNPs* and 16K *SNPs*, whatever the selection method (genetic or physical map), the imputation accuracy is very similar.

Introduction

The imputation is a robust tool able to infer the genotypes at an un-genotyped locus and has been adopted so as to minimize costs of genotyping in livestock breeding including Sheep. Since 2015 a new low-density array of about 16,000 *SNPs* is available in Sheep and its imputation accuracy on the 50K *SNPs* array was tested and confirmed as really performing in different French dairy Sheep breeds (Larroque *et al*., In Prep). Traditionally, the low-density arrays were created by degrading medium (or high) density *SNPs* arrays. The conserved *SNPs* were chosen in function of the physical distance, in order to optimize the spacing between markers and to obtain *SNPs* evenly spaced across the genome (Zhang and Druet, 2010, Dassonneville *et al.*, 2012). However, with this method, the chromosome ends have often an insufficient density, which could complicate the imputation in these regions. Therefore, it was demonstrated that if there was a quadrupled *SNP* density at the chromosomes ends, the imputation accuracy was higher and less variable (Bolormaa *et al.*, 2015).

The improvement of imputation in highly recombinant regions, as the chromosome ends, could also be possible with the use of the genetic distance instead of the physical distance. In fact, the genetic distance, computed in Morgan or centi-Morgan, used the recombination rate and so, gave more information for the imputation of genotypes. Actually, in livestock breeding, the genetic distance is not used for the construction of low-density arrays. On the other hands, in plants, and in particular in fruit trees, as apple tree or cherry tree, the arrays are created by combining genetic and physical location (Chagné *et al.*, 2012, Peace *et al.*, 2012). This *SNPs* selection enabled efficient genome saturation, helped to avoid redundancy, spanned gaps that would have occurred if the combination of the two maps were not used and finally led to higher

imputation accuracy.

In this work, we create different low-density *SNPs* sets, with 16,000 *SNPs* (16K), 10,000 *SNPs* (10K), 3,000 *SNPs* (3K) and 300 *SNPs* (LD) respectively. Because a genetic map is now available for the Sheep (Petit *et al.*, 2017), we created two sets for each case: one based on the physical distances and another based on genetic distance. We then compare by simulations the imputation accuracy for these two different ways of sets creation, in order to see if the genetic distance could allow to have best imputation accuracy and so to create really low-density arrays. These last could allow to impute for a very low cost.

Materials and Methods

Study population

The population used for the imputation simulations comes from the Lacaune breed sheep. We exploited a large population of 8,085 related rams genotyped with the medium density Illumina Ovine Beadchip® including 54,241 *SNPs* (50K array). Data were cleaning with the same criteria as those chosen by Larroque *et al*. (In Prep.); a call freq better than 0.97, a test for Hardy Weinberg Equilibrium was calculated and individuals which were better than 24 for a Khi^2 with 1 degree of freedom and a p-value < 10^{-5} were kept. Finally, a correction for the filiation was done using mandelian informations. After these quality controls, a total of 5,864 animals were kept. The remaining animals were separated in two groups: 4,718 Lacaune composed the training population and 1,146 were used as a validation set. The animals of the validation set were chosen in function of their birth date. In fact, the animals chosen for the validation set were all born in 2012 or after. Furthermore, they were used for a genomic cross-validation, because they already had a lot of daughters. All the remaining animals, with a previous vintage composed the training set.

Creation of low-density SNPs sets

In a precedent study, a genetic recombination map was created for the ovine Lacaune breed (Petit *et al.*, 2017). Briefly, 46,813 *SNPs* (50K array) were selected on the medium density Illumina

Ovine Beadchip® and after detecting crossovers (*COs*), which occurred between these *SNPs*, a family-based recombination rate (in cM/Mb) was estimated between the marker intervals. Thanks to this recombination rate, it was possible to obtain a map with the genetic distances (in cM) for each *SNP*.

We then used this map to create 4 *SNPs* sets with different densities thanks to a script written with the R software. We created two different sets: one with the physical distances and one with the genetic distances. They were then compared on their imputation quality.

We created 8 different low-density *SNPs* sets (4 different densities for each *SNPs* selection methods based either on physical map or on genetic map): a first with about 16,000 *SNPs*, a second with about 10,000 *SNPs*, a third with about 3,000 *SNPs* and the last one with only 300 *SNPs*. In order to create these sets, we first computed for each 46,813 *SNPs* of the 50K array the *MAFs* (Minor Allele Frequencies). To select the *SNPs* composing each low-density *SNPs* sets, we had to choose a method to select *SNPs* based on physical or genetic map. To select the *SNPs*, the genome was divided into equidistant segments (segment sizes were determined using either the genetic of the physical map). The number of segments was equal to the number of *SNPs* in the set (300, 3K, 10K or 16K), then, one *SNP* was selected by segment: the *SNP* has to be located in the 10% of the segment beginning and has to have the highest *MAF* in the region.

Finally, we obtained 8 low-density *SNPs* sets: 4 selected with a physical distance step, the *BOPM* markers (Based on Physical Map) sets and 4 selected with a genetic distance step, the *BOGM* markers (Based on Genetic Map) sets.

The compute of the total physical size of the 40K array gave a result of 2,438,873,539 base pairs and the array had a genetic size of 3,590.621 cM. These two sizes were used to determine the different steps of the *SNPs* choice for the 300*SNPs*, the 3K, the 10K and the 16K sets. The markers were selected every 8,129,578 bp, 812,958 bp, 243,887 bp and 152,430 bp respectively for the sets selected on the physical size, and every 11.97 cM, 1.20 cM, 0.36 cM and 0.22 cM respectively for the sets selected on the genetic distance. For each of these steps, we choose the closet *SNP* that had the highest *MAF*.

Finally, we obtained 2 sets with 341 *SNPs* (*BOPM* set) and 340 *SNPs* (*BOGM* set), 2 sets with 3,033 *SNPs* (*BOPM* set) and 3,034 *SNPs* (*BOGM* set), 2 sets with 9,557 *SNPs* (*BOPM* set) and 9,857 *SNPs* (*BOGM* set) and 2 last sets with 15,696 *SNPs* (*BOPM* set) and 15,599 (*BOGM* set).

Imputation of the low-density SNPs sets on the 50K SNPs array

We created 8 low-density *SNPs* sets and we wanted to impute them on the 50K array. We used the pipeline created by Larroque *et al*. (In Prep) to realize this imputation.

There were different steps in this pipeline:

- Two populations; the training population and the validation population were created based on the 5,864 animals genotyped with the 50K *SNPs* array.
- We kept only the genotypes for the selected *BOPM* or *BOGM SNPs* for the validation population.
- We used the FIMPUTE 2.2 (Sargolzaei *et al.*, 2014) software for the imputation of the unknown *SNPs* of the validation population using the 50K genotyping of the training population.
- We computed for each chromosome and each imputed *SNP*, the concordance rate (*CR*), determined as the proportion of correctly imputed markers out of all markers that were inferred after imputation, the error rate (computed as 1 – *CR*) and the allelic squared Pearson correlation (r^2).

Test of the imputation quality for the 8 low-density SNPs sets

The chromosomes were split into three pieces: the beginning of the chromosome, the middle of the chromosome and the end of the chromosome. The beginning and the end were chosen using the Sheep recombination of Petit *et al.* (2017); because the recombination rate was the most elevated at the extremities of the chromosomes, limits of the beginning and of the end were considered as the inflection points of the recombination rate. Metacentric chromosomes, the first 3 Sheep chromosomes, had a very low recombination rate at the centromere (Petit *et al.*, 2017), that is why the centromeric regions were discarded.

The imputation accuracy was essentially studied by comparing the poorly imputed markers using the *CR*. We considered that a marker is badly imputed when its *CR* was fewer than 98% (Dassonneville *et al.*, 2012). For each set, the proportion of poorly imputed markers, it meant the number of poorly imputed markers among all the imputed *SNPs* selected *BOGM* or *BOPM*, of each part of the chromosome (middle and extremities, which corresponded to the sum of the beginning and the end) was computed. Khi^2 tests allowed to indicate if there was a significant difference between these proportion.

Results

Test of the imputation quality for the 8 low-density SNPs sets

Markers distribution and imputation accuracy were studied for the different sets and for each part of the chromosomes. In each case, we plotted the proportion of poorly imputed markers and we added the distribution of not imputed markers selected *BOPM* or *BOGM*.

We analyzed the beginning and the end of each chromosome separately and together and we observed that the results were similar, thus for the further analysis, we only observed the middle and the extremities of the chromosomes (beginning and end of the chromosome together).

First, we observed the imputation accuracy for the chromosomes extremities for each *SNPs* set. The 300 *SNPs* set was really to small and no significant differences could be highlighted between markers selected *BOPM* or *BOGM*. An array was developed for the kinship assignment in Sheep (Tortereau *et al.*, 2014). It contained 249 *SNPs* selected to be informative for 30 French Sheep breeds. We thus compared the imputation accuracy of this existing array with markers selected *BOPM*, however, no significant differences were detected. The three sets were completely similar (**see Figure 1**). The number of selected markers was really too small to be plotted on a figure with these sets.

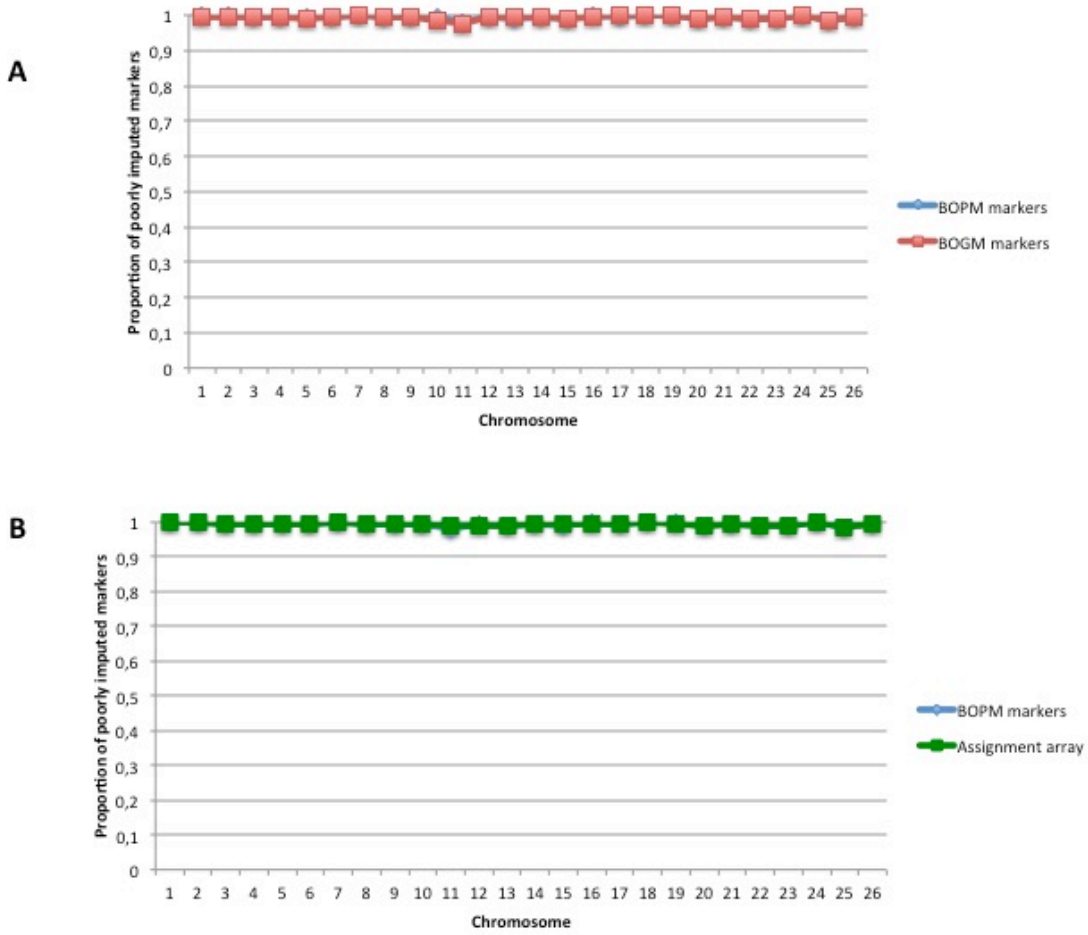

Figure 1: Proportion of poorly imputed markers at the extremities of the chromosomes for the 300 SNPs selected BOPM and the 300 SNPs selected BOGM (A) and for the 300 SNPs selected BOPM and the assignment array (B)

The proportion of poorly imputed markers for each of the three datasets and at each chromosome was reported.

For the 3K *SNPs* sets, some significant differences were observed between the markers selected *BOPM* or *BOGM* (**see Figure 2**); markers selected *BOGM* had globally the least of poorly imputed markers.

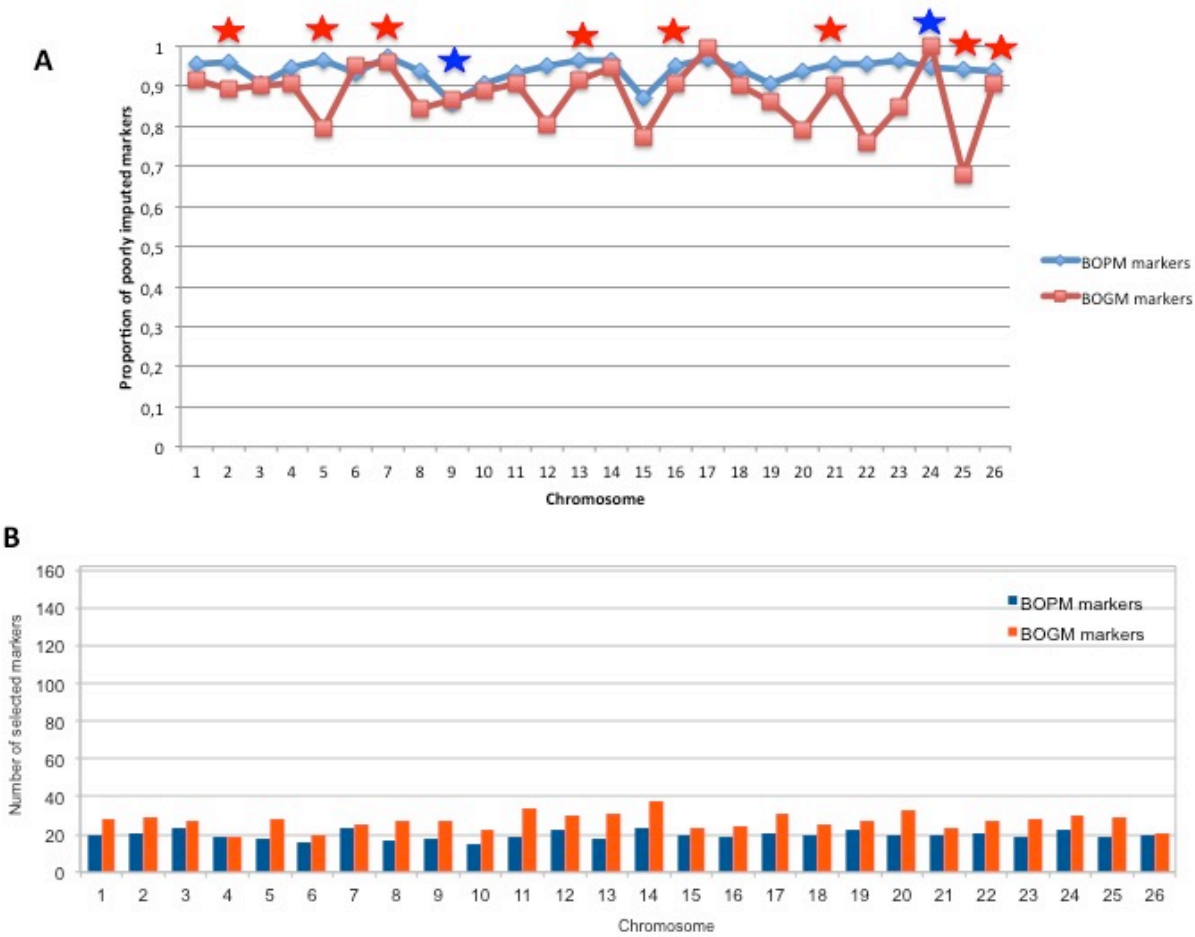

Figure 2: Proportion of poorly imputed markers (A) and distribution of selected markers (B) at the extremities of the chromosomes for the 3K SNPs selected BOPM and the 3K SNPs selected BOGM

A/ The proportion of poorly imputed markers for each of the two datasets was reported and compared using a Khi^2 test. The stars indicated a significant difference between the two datasets. The red stars indicated that markers selected *BOGM* had a better imputation accuracy, on contrary, the blue stars indicated that markers selected *BOPM* had a better imputation accuracy. **B/** Distribution of selected marker (non imputed markers) for the two datasets.

Especially, the markers selected *BOGM* were more numerous at the extremities than the markers selected *BOPM*, up to about 40% more. It appeared that, when there were more *SNPs* selected *BOGM*, the imputation accuracy was also better, for example for the chromosomes 5 or 20. The biggest significant differences between markers selected *BOPM* or *BOGM* were observed for the 10K *SNPs* set (**see Figure 3**).

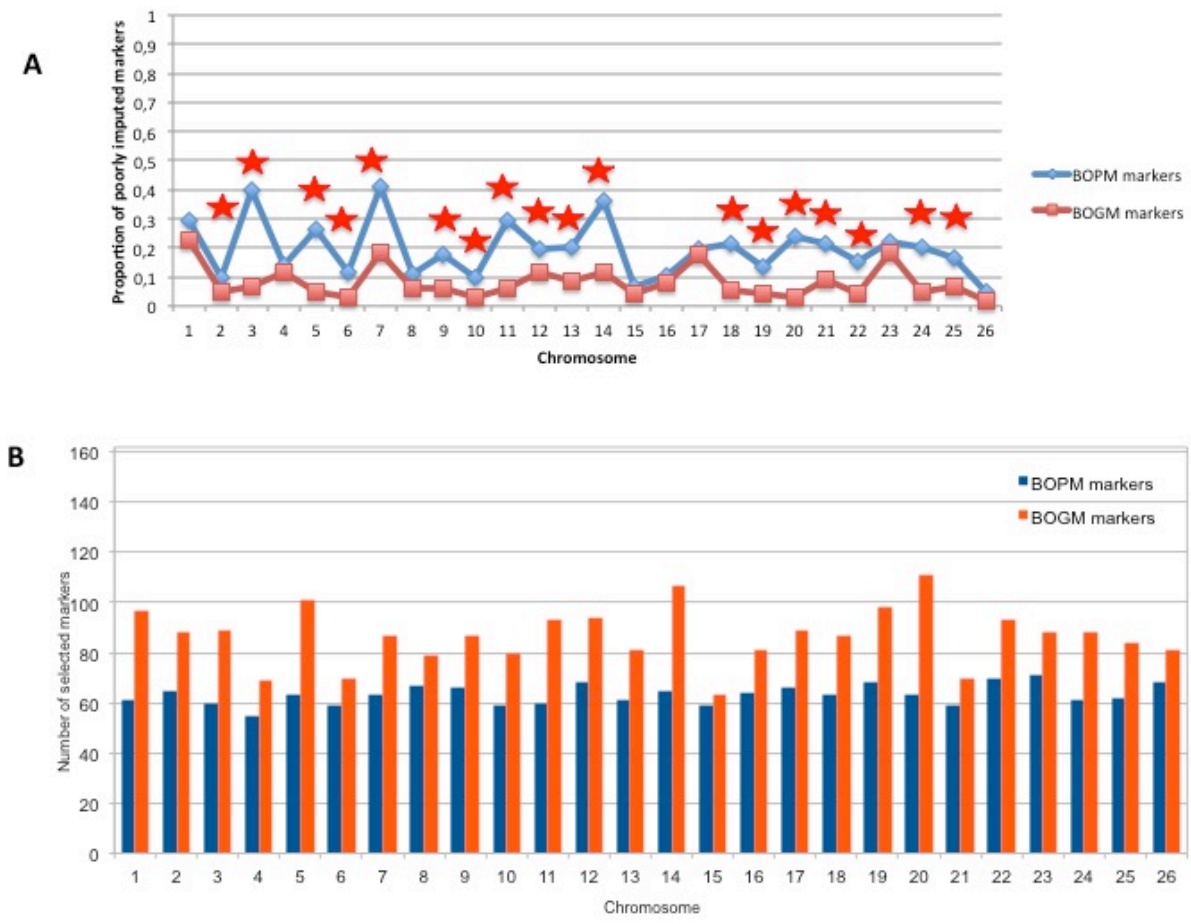

Figure 3: Proportion of poorly imputed markers (A) and distribution of selected markers (B) at the extremities of the chromosomes for the 10K SNPs selected BOPM set and the 10K SNPs selected BOGM

A/ The proportion of poorly imputed markers for each of the two datasets was reported and compared using a Khi^2 test. The stars indicated a significant difference between the two datasets. The red stars indicated that markers selected *BOGM* had a better imputation accuracy, on contrary, the blue stars indicated that markers selected *BOPM* had a better imputation accuracy. **B/** Distribution of selected marker (non imputed markers) for the two datasets.

There was until 55% more markers selected *BOGM* and the relation between the number of markers and the imputation accuracy was really observable, especially for the chromosomes 11, 14 or 20. With the 16K *SNPs* sets, there were globally few significant differences between markers selected *BOPM* and markers selected *BOGM* set: only 9 chromosomes showed significant differences (**see Figure 4**).

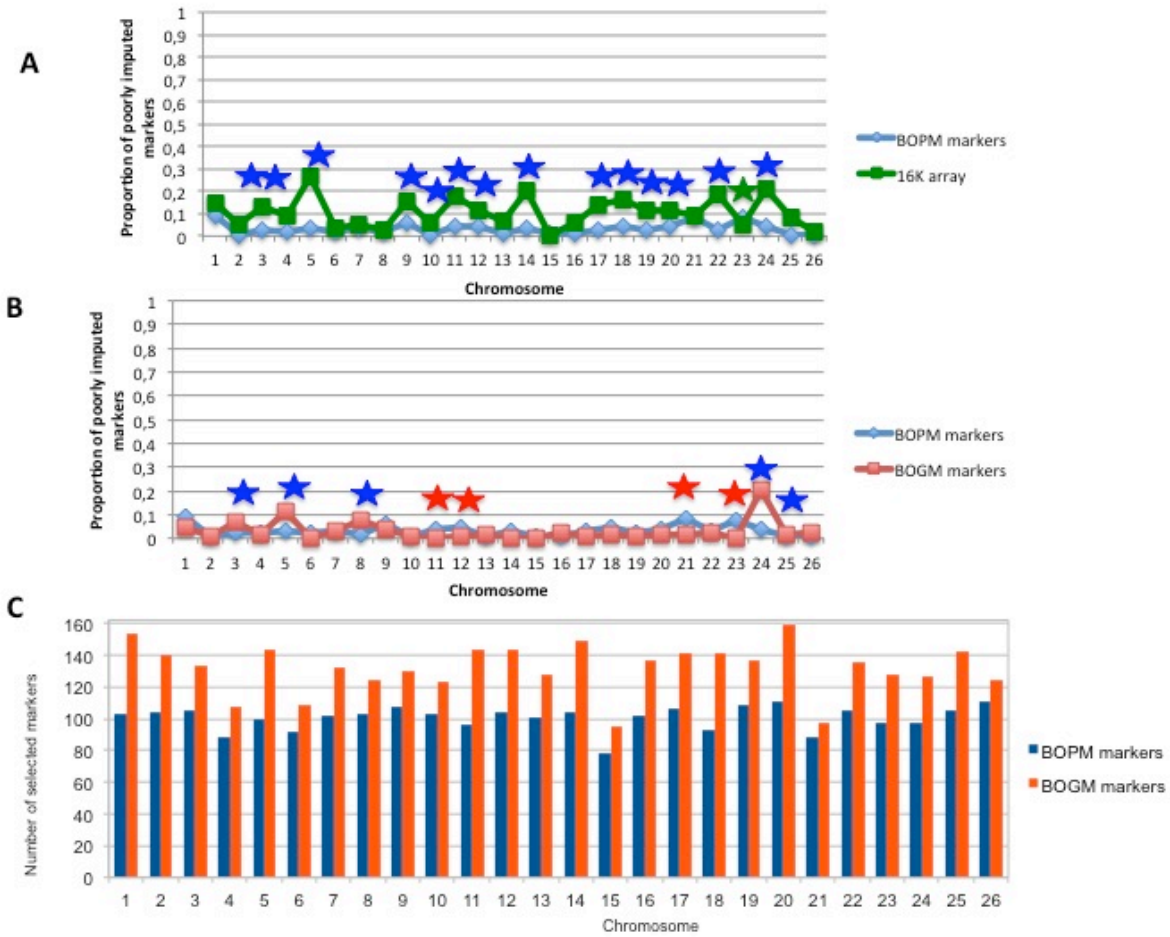

Figure 4: Proportion of poorly imputed markers (A) and distribution of selected markers (B) at the extremities of the chromosomes for the 16K SNPs selected BOPM set and the 16K SNPs selected BOGM and proportion of poorly imputed markers for the 16K SNPs selected BOPM and the existing 16K array (C)

A/ The proportion of poorly imputed markers for markers selected *BOPM* and markers selected *BOGM* was reported and compared using a Khi^2 test. The green star indicated a significant difference between the two datasets. **B/** The proportion of poorly imputed markers for markers selected *BOPM* and the existing 16K array was reported and compared using a Khi^2 test. The stars indicated a significant difference between the two datasets. The red stars indicated that markers selected *BOGM* had a better imputation accuracy, on contrary, the blue stars indicated that markers selected *BOPM* had a better imputation accuracy and the green stars indicated that markers of the existing 16K had a better imputation accuracy. **C/** Distribution of selected marker (non imputed markers) for markers selected *BOPM* or *BOGM*.

As for the three other *SNPs* sets, there were more markers selected *BOGM*, up to about 30%. As for the 300 *SNPs* sets, there was already an existing 16K array, which was developed to impute animals for low costs. Markers were thus soundly chosen and so we expected to that this array had a better imputation accuracy than markers selected *BOPM*. However, 16 chromosomes

of the existing 16K array presented a significant difference with markers selected *BOPM* and were more poorly imputed.

We made the same comparisons for the middle of the chromosomes. As for the chromosomes extremities, no significant differences were detected between markers selected *BOPM* and markers selected *BOGM* for the 300 *SNPs* set. In the same way, the assignment array and markers selected *BOPM* were completely similar (**see Figure 5**).

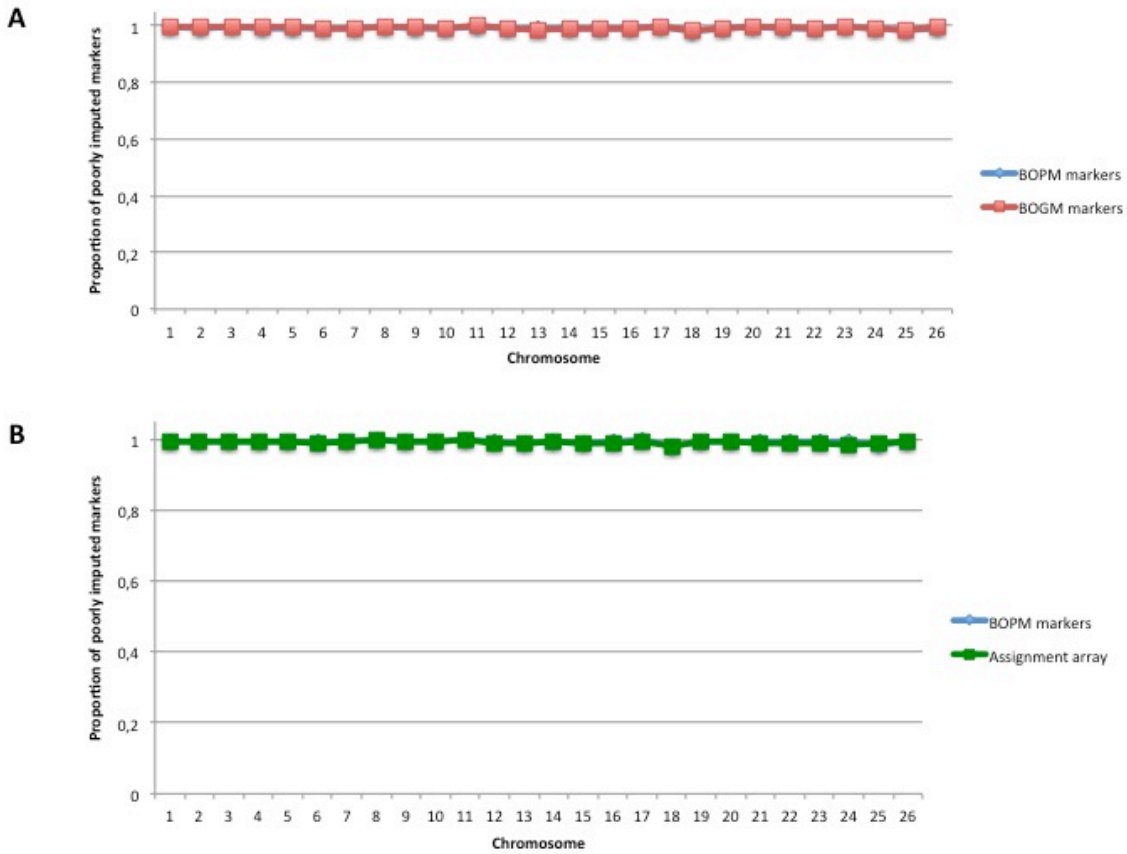

Figure 5: Proportion of poorly imputed markers on the middle of the chromosomes for the 300 SNPs selected BOPM and the 300 SNPs selected BOGM (A) and for the 300 SNPs selected BOPM and the assignment array (B)

The proportion of poorly imputed markers for each of the three datasets and at each chromosome was reported.

Contrary to the extremities of the chromosomes, there were more markers selected *BOPM* in the middles of the chromosomes. The chromosome 6 presented about 40% more markers selected *BOPM*. For the 3K *SNPs* sets, nearly all the chromosomes had a significant difference;

markers selected *BOGM* having a less good imputation accuracy (**see Figure 6**).

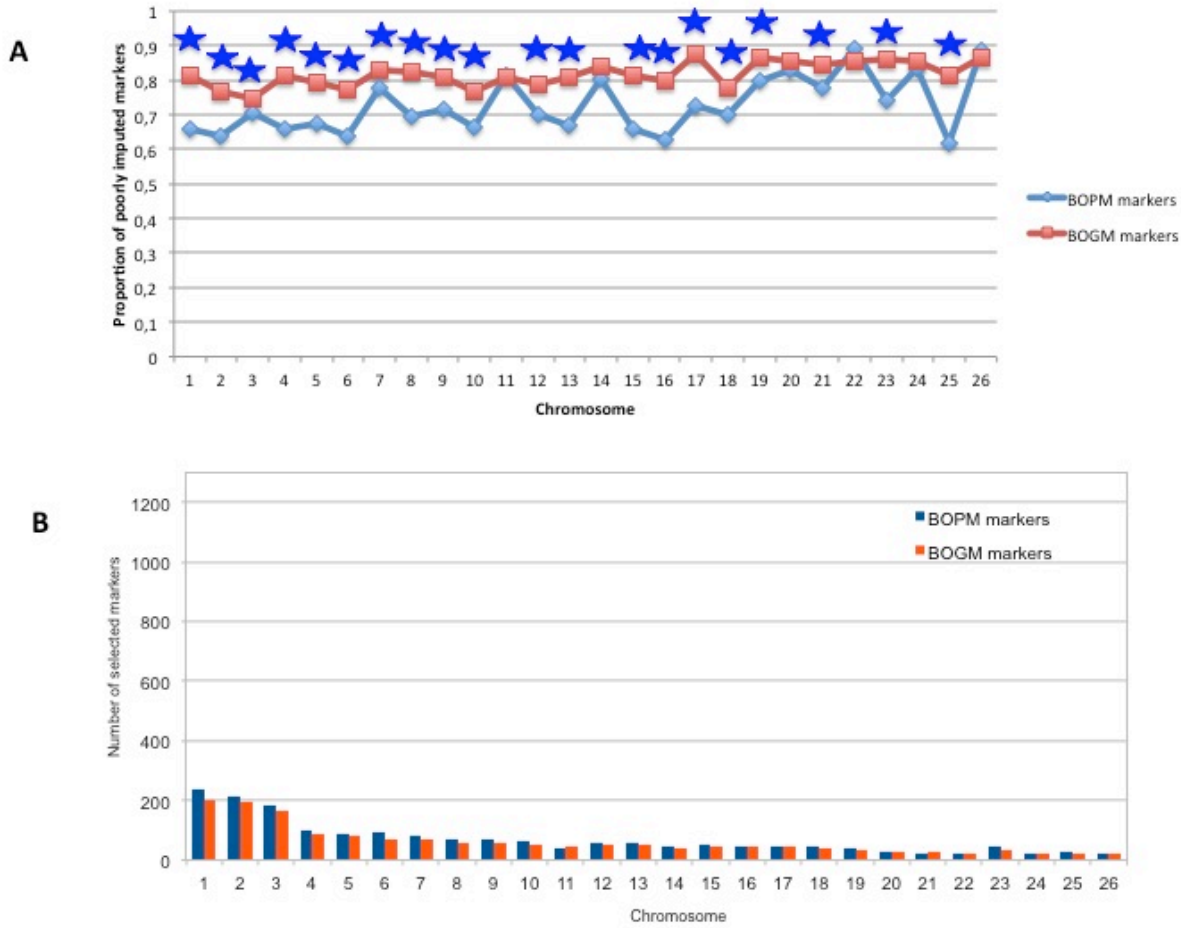

Figure 6: Proportion of poorly imputed markers (A) and distribution of selected markers (B) in the middle of the chromosomes for the 3K SNPs selected BOPM and the 3K SNPs selected BOGM

A/ The proportion of poorly imputed markers for each of the two datasets was reported and compared using a Khi^2 test. The stars indicated a significant difference between the two datasets. The red stars indicated that markers selected *BOGM* had a better imputation accuracy, on contrary, the blue stars indicated that markers selected *BOPM* had a better imputation accuracy. **B/** Distribution of selected marker (non imputed markers) for the two datasets.

Markers selected *BOPM* were in majority for 20 chromosomes, which could explain why the imputation accuracy was better. With the 10K *SNPs* sets, 16 chromosomes had a significant difference between markers selected *BOPM* or *BOGM*; markers selected *BOPM* having a better imputation accuracy (**see Figure 7**).

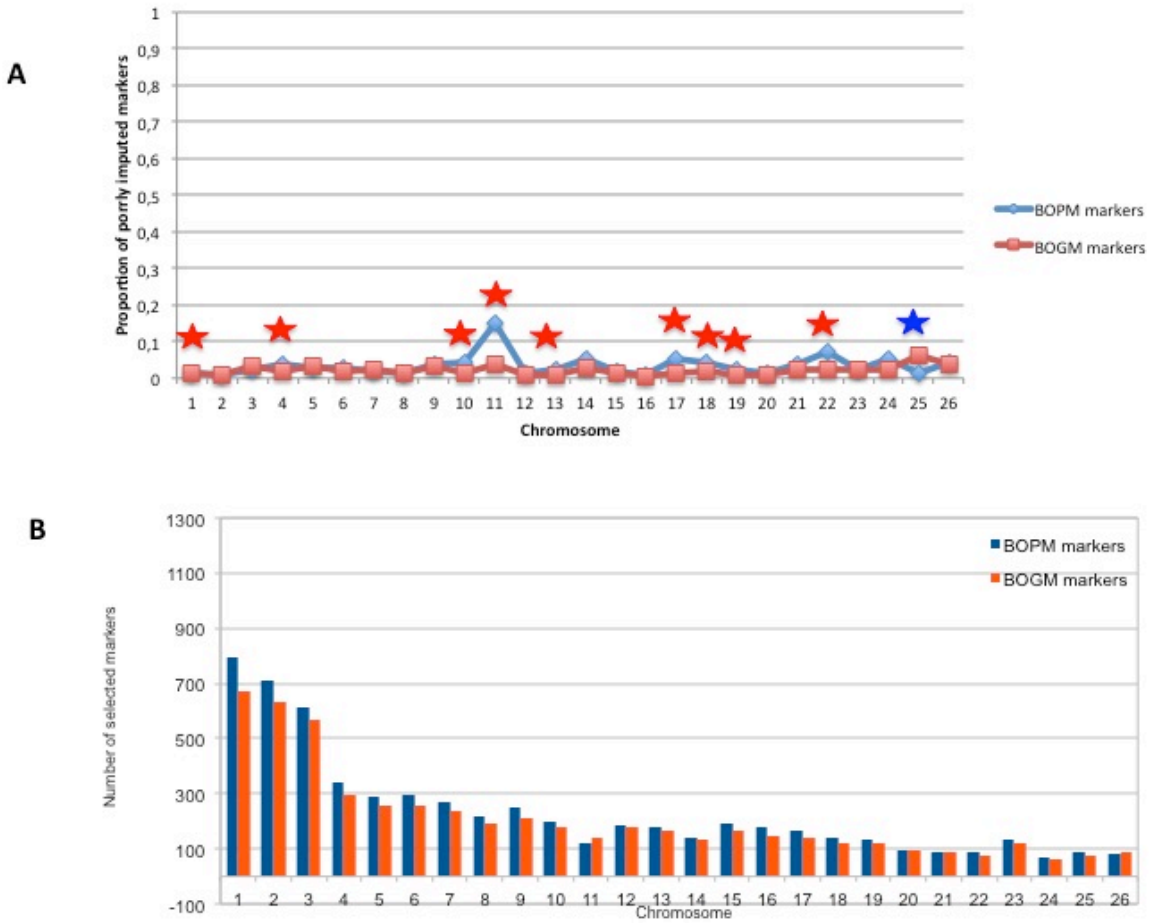

Figure 7: Proportion of poorly imputed markers (A) and distribution of selected markers (B) in the middle of the chromosomes for the 10K SNPs selected BOPM set and the 10K SNPs selected BOGM

A/ The proportion of poorly imputed markers for each of the two datasets was reported and compared using a Khi^2 test. The stars indicated a significant difference between the two datasets. The red stars indicated that markers selected *BOGM* had a better imputation accuracy, on contrary, the blue stars indicated that markers selected *BOPM* had a better imputation accuracy. **B/** Distribution of selected marker (non imputed markers) for the two datasets.

Twenty chromosomes had more markers selected *BOPM*, up to about 20% more for the chromosomes 1, 9 or 16. Finally, for the 16K *SNPs* sets, there were globally few significant differences between markers selected *BOPM* or *BOGM* (**see Figure 8**).

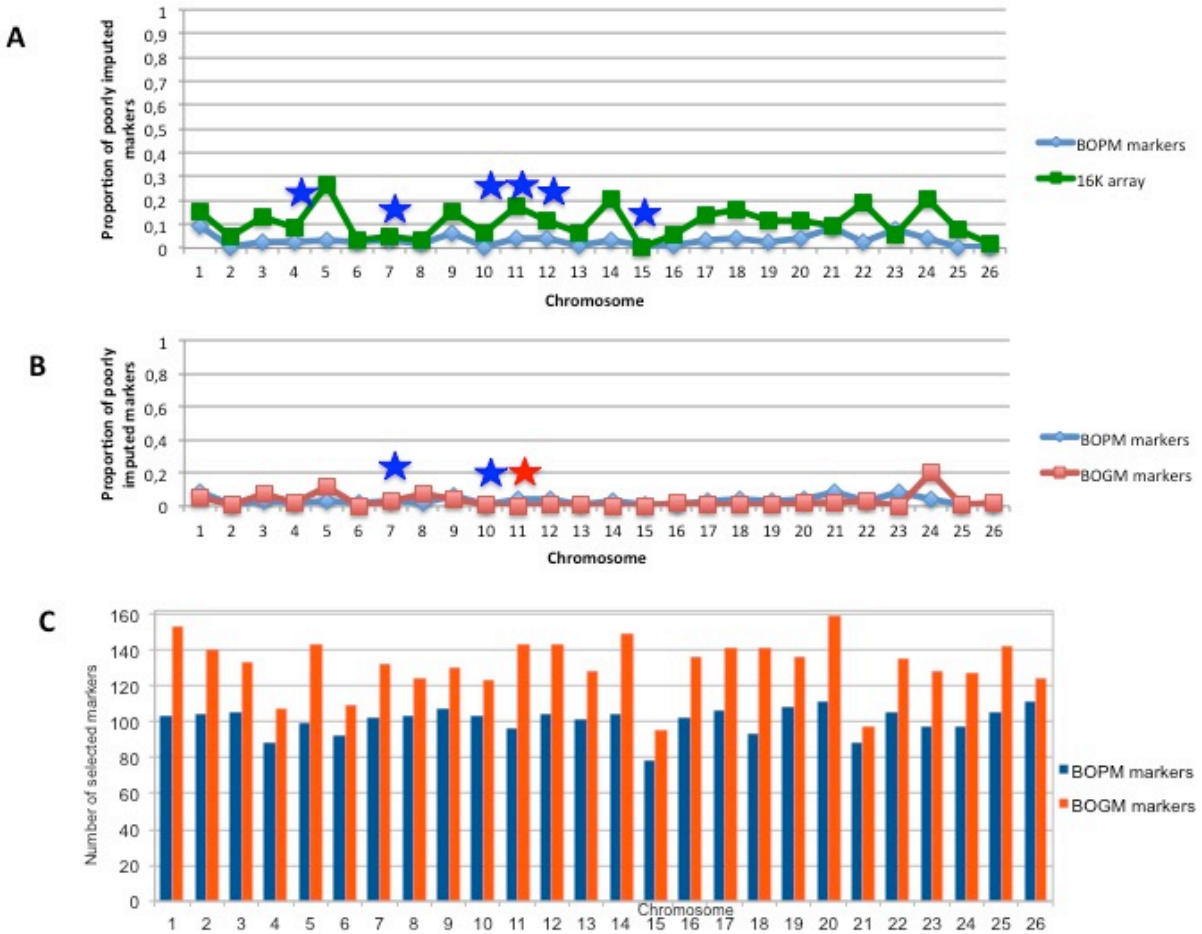

Figure 8: Proportion of poorly imputed markers (A) and distribution of selected markers (B) in the middle of the chromosomes for the 16K SNPs selected BOPM set and the 16K SNPs selected BOGM and proportion of poorly imputed markers for the 16K SNPs selected BOPM and the existing 16K array (C)

A/ The proportion of poorly imputed markers for the *BOPM* set and the *BOGM* set was reported and compared using a Khi^2 test. The green star indicated a significant difference between the two datasets. **B/** The proportion of poorly imputed markers for markers selected *BOPM* and the existing 16K array was reported and compared using a Khi^2 test. The stars indicated a significant difference between the two datasets. The red stars indicated that markers selected *BOGM* had a better imputation accuracy, on contrary, the blue stars indicated that markers selected *BOPM* had a better imputation accuracy and the green stars indicated that markers of the existing 16K had a better imputation accuracy. **C/** Distribution of selected marker (non imputed markers) for markers selected *BOPM* or *BOGM*.

Furthermore, the existing 16K array and markers selected *BOPM* were really similar; only 6 chromosomes presented a significant difference. As for the three others sets, the number of markers selected *BOPM* was greater: between 10 and 20% more.

10K *SNPs* selected *BOGM* showed the greater interest for the creation of a low-density

array. In fact, there was a clear improvement at the extremities of the chromosomes, with sometimes 40% more markers. It could thus be interesting to study the economic impact of the creation of this array, compared to the 16K array already existing. In fact, create a new array is really expensive, but it is possible to add custom markers on the already existing 16K array. Thus, using of markers selected *BOGM* could allow to choose specific markers, especially at the extremities of the chromosomes, which could improve the imputation accuracy.

Interestingly, the already existing 16K array showed a less good imputation accuracy than the 16K *SNPs* selected *BOPM*. This array was created combining two types of data: genomic selection data with about 12,000 markers and 4,000 markers, which came from an assignment array. The markers were chosen in function of their physical distance and in function of their *MAF*. However, these markers and *MAF*s are not specific of the Lacaune breed; they were chosen in a lot of international different Sheep breeds. Thus, the existing 16K array is not specific for the Lacaune breed, contrary to markers selected *BOPM*, for which chosen markers and *MAFs* were specific of the Lacaune. Because the existing 16K is non-specific, this could explain why the imputation accuracy is weaker with this array. In order to confirm (or infirm) this hypothesis, it could be interesting to create a new 16K markers selected *BOPM*, where the *MAFs* of different populations (such as the 71 breeds from the Sheep HapMap project) are used to choose the markers. The differences between the 16K *SNPs* selected *BOPM* and the existing 16K array may be lesser.

We noticed that the number of supplementary markers was always more important at the extremities of the chromosomes, for markers selected *BOGM*, that for the middle of the chromosomes. Furthermore, markers selected *BOGM* showed greater imputation accuracy at the extremities. This potential link seemed thus indicate that enrich the extremities with markers selected *BOGM* allowed to increase the imputation accuracy. The chromosomes extremities had high recombination rates, so poor linkage disequilibrium; yet, good linkage disequilibrium is necessary to impute correctly. Densify the extremities with markers selected *BOGM* could thus compensate for this lost of linkage disequilibrium. To study this hypothesis, it would be interesting to create *SNPs* sets using the genetic distance and the linkage disequilibrium information.

Discussion

Height low-density *SNPs* sets were created with two different ways: based on physical distances or based on genetic distances. The imputation accuracy and the markers distribution were compared between the two ways of creation.

The 300 *SNPs* set was really to small and, even if we used the genetic distances, we could not increase the imputation accuracy, but it was however not decreased and markers selected *BOPM* or *BOGM* were completely similar. For the 3K *SNPs* set, the number of markers is also weak and the imputation accuracy was not increased in the middle of the chromosomes. However, the number of markers selected *BOGM* was greater at the extremities of the chromosomes, which could allow a better imputation accuracy. This set could thus be interesting for the imputation if we created an array combining markers selected *BOPM* and *BOGM*. For better results, it could also be relevant to study a 5K set, which would have a few more markers. There was not a lot of gain with 16K *SNPs* selected *BOPM*, probably because there were already enough markers and so to add markers selected *BOPM* was maybe not really relevant.

Petit *et al.* (2017) noticed that the Lacaune and another Sheep breed, the Soay, had very similar recombination map, since the two populations were very distinct. This result could highlight that we could maybe use the Lacaune recombination map to create low-density *SNPs* sets in different Sheep breeds, which would be a gain of time. Using thus map allowed to compare the imputation accuracy using different *SNPs* sets, as here, but also to compare the imputation accuracy between the breeds.

Conclusion

The aim of this study was to compare the imputation accuracy of different low-density *SNPs* sets, created either using physical distances, or using genetic distances. Height *SNPs* were thus created: two for a 300 *SNPs* set, two for a 3K *SNPs* set, two for a 10K *SNPs* set and two others for a 16K *SNPs* set. The comparison between the sets revealed that the most interesting is the 10K

SNPs set, where markers selected *BOGM* has always a better imputation accuracy than the markers selected *BOPM*, which could be due to a more important number of markers selected *BOGM*. However, the existing 16K array had a worse imputation accuracy than the 16K markers selected *BOPM*, which could be explained by the non-specificity of this array for the Lacaune breed.

Literature cited

- Bolormaa S., Gore K., Van der Werf JHJ., Hayes BJ. and Daetwyler HD. (2015). Design of a low-density *SNP* chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. *Animal Genetics*, 46, pp. 544-556.
- Dassonneville R., Fritz S., Ducrocq V. and Boichard D. (2012). Imputation performances of 3 low-density marker panels in beef and dairy cattle. *Journal of Dairy Science*, 95, pp. 4136-4140.
- Chagné D., Crowhurst RN., Troggio M., Davey MW., Gilmore B. *et al.* (2012). Genome-wide *SNP* detection, validation and development of an 8K *SNP* array for apple. *PLoS ONE* 7(2): e31745.
- Peace C., Bassil N., Main D., Ficklin S., Rosyara UR. *et al.* (2012). Development and evaluation of a genome-wide 6K *SNP* array for diploid sweet cherry and tetraploid sour cherry. *PLoS ONE* 7(12): e48305.
- Petit M., Astruc JM., Sarry J., Drouilhet L., Fabre S. *et al.* (2017). Variation in recombination rate and its genetic determinism in sheep (*Ovis Aries*) populations from combining multiple genome-wide datasets. *bioRxiv.*
- Sargolzaei M., Chesnais JP. and Schenkel FS. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, 15.
- Tortereau F., Palhière I., Moreno C., Tosser-Klopp G., Barbotte L. *et al.* (2014). Assignation de parentés pour les populations françaises de petits ruminants en sélection. *Rencontres autour des Recherches sur les Ruminants*, 21, pp. 257-260.
- Zhang Z. and Druet T. (2010). Marker imputation with low-density marker panels in Dutch Holstein cattle. *Journal of Dairy Science*, 93, pp. 5487-5494.

.

Discussion Générale

Discussion générale

L'étude de la recombinaison chez les animaux d'élevage pourrait servir de modèle pour l'Homme. En effet, l'étude des phénotypes de recombinaison nécessite, en particulier pour le taux de recombinaison individuel, un nombre important de méioses individuelles, afin de pouvoir repérer suffisamment de crossing-overs, lorsqu'il n'est pas possible d'avoir accès directement à des méthodes plus précises, telles que le « sperm-typing ». Cela nécessite donc d'avoir des parents avec beaucoup de descendants, donc des grandes familles. Or, il est parfois difficile d'obtenir de tels échantillons chez l'humain. Le plus grand « pédigrée » existant chez l'Homme à ce jour provient de la population Islandaise, dans lequel 23 066 individus ont été génotypés, ce qui a permis d'obtenir 14 140 méioses (Kong *et al.*, 2004). En revanche, chez les animaux d'élevage, c'est assez souvent que des pères ont une centaine de descendants ; ainsi dans l'étude de Ma et collaborateurs (2015), le pédigrée à disposition contenait 185 917 familles de trois générations, ce qui a permis d'obtenir 132 331 méioses. De même, une autre étude a pu obtenir 119 848 méioses provenant de 14 401 individus Holstein (Kadri *et al.*, 2016). Un aussi grand nombre de données permet d'obtenir un phénotype plus précis. La disponibilité d'autant d'individus ayant phénotypes et génotypes permet des identifications plus précises et plus robustes de gènes candidats lors de la recherche de *QTLs*. Au vu des résultats présentés dans la littérature, ce serait sûrement l'espèce porcine qui serait la plus intéressante pour servir de modèle à l'étude de la recombinaison chez l'Homme. En effet, leur taux de recombinaison sont similaires ; proches de 1 cM/Mb (Chowdhury *et al.*, 2009, Tortereau *et al.*, 2012), ils ont des patrons de recombinaison très proches, une variation de la recombinaison du même ordre de grandeur et de plus, contrairement aux ovins ou aux bovins, ce sont les femelles qui recombinent plus que les mâles, comme chez l'Homme. En revanche, le déterminisme génétique des phénotypes de recombinaison n'a pas encore été étudié chez le porc, laissant donc la possibilité que les gènes candidats soient différents de ceux identifiés chez l'Homme.

Dans cette thèse, des cartes de recombinaison de haute densité, pour la recombinaison méiotique et pour la recombinaison historique, ont été créées, ce qui a permis de caractériser la recombinaison chez la Lacaune et d'identifier des points chauds historiques. Cependant, il n'a pas été possible d'étudier le deuxième phénotype de recombinaison : le biais d'usage des points

chauds.

I. Amélioration des cartes de recombinaison

I. 1. *Amélioration des cartes de recombinaison méiotique*

I. 1. a. *La combinaison de différents jeux de données*

L'amélioration des cartes de recombinaison méiotiques pourrait être permise en combinant plusieurs jeux de données issus de différentes races. En effet, nous avons montré précédemment que la combinaison des cartes de recombinaison issues des races Lacaune et Soay permettait d'améliorer la précision des cartes de recombinaison méiotique, puisque les deux races ont des patrons de recombinaison quasiment identiques. La race ovine Manech Tête Rousse a commencé à collecter de nombreuses données à la suite du programme « Genomia » (2010-2012) qui permettait de réfléchir à la mise en place de la sélection génomique dans cette race. Il serait donc intéressant de créer des cartes de recombinaison dans cette race afin de savoir si la distribution des taux de recombinaison est également conservée dans cette population et si elle peut donc être utilisée pour améliorer encore plus les cartes de recombinaison. Il serait également intéressant d'observer si le déterminisme génétique du taux de recombinaison individuel est similaire à celui des Lacaune ou des Soay ou s'il est différent. De même, le projet *Romane Ite Domum* permettant de génotyper 30 familles avec la puce 600K, pourra également être utilisé pour améliorer les cartes, notamment aux extrémités (à moins de 4 Mb), où la détection des crossing-overs est très peu précise. D'autres races ovines présentant suffisamment de données pourront également, à terme, être utilisées pour la création de ces cartes.

I. 1. b. *Création de cartes de recombinaison âge- et sexe-spécifiques*

Il pourrait également être intéressant d'étudier un potentiel effet de l'âge sur la recombinaison. En effet, il a été montré chez l'Homme (en particulier chez la femme), que l'âge impactait la recombinaison : les descendants de mères de plus de 35 ans présentaient ainsi un taux de recombinaison plus élevé (Coop *et al.*, 2008). Cet effet de l'âge maternel pourrait être dû à

une sélection contre les oocytes qui ne présentent pas suffisamment de recombinaisons, afin d'éviter l'accumulation de mutations délétères. Bien que ce résultat n'ait pas été démontré chez l'homme, il serait intéressant de savoir s'il peut se retrouver chez les béliers. Il est vrai qu'en élevage, les béliers sont abattus plutôt jeunes par rapport à leur espérance réelle de vie (autour de 6 ans, alors qu'ils peuvent vivre près de 13 ans), cependant, si nous obtenons un nombre suffisant d'individus jeunes (moins de 2 ans), d'individus moyens (entre 2 et 4 ans) et d'individus « âgés » (plus de 4 ans), il serait peut-être possible d'étudier l'effet de l'âge sur le taux de recombinaison et ainsi de créer des cartes de recombinaison propres à chaque catégorie d'âge.

Chez certaines espèces, il a été montré qu'il pouvait y avoir un effet du sexe sur le taux de recombinaison ; les mâles et les femelles ayant des taux différents. C'est par exemple le cas chez l'Homme, où un taux de recombinaison de 1,25 cM/Mb est observé chez la femme et un taux de 0,81 cM/Mb chez l'homme (Chowdhury et *al.*, 2009). C'est également le cas chez le porc, où, à l'échelle du génome, les femelles ont un taux de recombinaison plus élevé que les mâles (Tortereau et *al.*, 2012). De même chez le chien, où les femelles recombinent 1,4 fois plus que les mâles (Neff et *al.*, 1999), ou chez le poisson zèbre, la femelle recombinant 2,7 fois plus que le mâle. Dans cette espèce, il semble que la recombinaison soit encore plus réprimée que chez l'Humain « mâle » (Singer *et al.*, 2002). Chez de nombreux Mammifères, ce serait donc les femelles qui auraient un taux de recombinaison plus élevé. Pour ce qui est de l'héritabilité du caractère, on observe également une différence entre les sexes : chez l'Homme, les femmes ayant une héritabilité plus élevée (Kong et *al.*, 2014).

En revanche, en bovin, c'est l'inverse, avec une longueur de la carte génétique plus grande pour le mâle que pour la femelle (Ma et *al.*, 2015). De même pour le marsupial wallaby, où les mâles ont une carte génétique plus longue que celle des femelles (Zenger et *al.*, 2002). Comme nous l'avons évoqué précédemment, c'est également vrai chez le mouton : les mâles ayant une carte de recombinaison plus longue que les femelles (Maddox et Cockett, 2008) et en race Soay (Johnston *et al.*, 2016), un effet sexe est également observé au niveau des cartes génétiques et du déterminisme génétique du taux de recombinaison individuel.

Chez les ovins Soay, les *QTLs* identifiés sont dans les mêmes régions, et bien que *RNF212* n'a pas été proposé comme gène candidat, il est localisé dans la région du *QTL*. Certains allèles de *RNF212* ont, chez l'Homme, des effets inverses : ils augmentent la recombinaison chez la femme,

mais la diminuent chez l'Homme (Kong *et al.*, 2008). Cela n'a pas pu être établi chez les Soay, puisque *RNF212* n'a été associé à la variation de la recombinaison que chez les femelles. En revanche, s'il est possible de faire un jour cette recherche de *QTLs* en femelles Lacaune, on pourra peut-être également mettre en évidence cet antagoniste, ou un autre résultat, dans la race Lacaune.

Cette différence de taux de recombinaison entre les mâles et les femelles est appelée « l'hétérochiasmie », elle concerne autant la fréquence moyenne de recombinaison que la distribution des évènements de recombinaison sur le génome.

Etant donné qu'une différence du taux de recombinaison existe entre les mâles et les femelles dans de nombreuses espèces, et plus particulièrement chez le mouton, il serait intéressant de savoir si c'est également le cas en race Lacaune. Malheureusement, très peu de femelles sont génotypées dans cette race et aucune ne faisait partie des 345 animaux pour lesquels nous avions établi un phénotype du taux de recombinaison individuel. A ce jour, il n'est donc pas possible de rechercher une différence entre les béliers et les brebis Lacaune. Pour ce faire, il faudrait disposer d'un nombre suffisant de femelles génotypées (effectif proche de celui existant chez les béliers Lacaune). Ces brebis doivent avoir également au moins 2 ou 4 descendants génotypés, selon que l'un de leur parent est génotypé ou non.

Nous pourrions ensuite comparer les moyennes des taux de recombinaison (en crossing-overs par méiose et en cM/Mb) entre les brebis et les béliers et voir si, comme pour les bovins et certains moutons, les béliers Lacaune recombinent plus également, cela permettra d'obtenir des cartes de recombinaison sexe-spécifiques. Dans les espèces de rente, une recombinaison plus importante chez les mâles pourrait être due à la sélection artificielle réalisée par l'Homme (Ma *et al.*, 2015). En effet, la sélection s'est majoritairement portée sur la voie mâle et ceci pourrait avoir augmenté la recombinaison si la sélection a un effet positif direct ou indirect sur la recombinaison.

I. 2. *Création de cartes de recombinaison haute résolution*

I. 2. a. *Utilisation de la recombinaison historique*

Nous avons vu qu'établir des cartes de recombinaison historiques, et ainsi détecter des

points chauds historiques demande d'avoir une très bonne résolution et donc des données très précises. En Lacaune, il faudrait soit augmenter le nombre d'individus génotypés sur la puce 600K, cependant cela reste très coûteux, ou alors, il faudrait pouvoir utiliser des données de séquence. Actuellement, il n'y a que 5 Lacaune qui sont intégralement séquencés, donc bien trop peu pour pouvoir être utilisés. En revanche, le projet Nextgen, qui avait pour but de développer des méthodologies optimisées pour la préservation de la biodiversité des animaux d'élevage, a permis de séquencer les génomes complets de 160 moutons marocains issus de localités représentatives de la diversité des conditions climatiques, écologiques et des systèmes d'élevage (Pompanon *et al.*, 2015). Il pourrait donc être intéressant d'utiliser ces données pour avoir avoir une très bonne résolution des points chauds historiques. Au cours de notre étude, il a été nécessaire de supprimer les portions du génome se trouvant à moins de 4 Mb de l'extrémité de chaque chromosome. En effet, nos données n'étaient pas suffisamment précises pour permettre d'estimer correctement les crossing-overs dans ces régions. L'utilisation de séquences pourrait donc permettre d'améliorer la prédiction de la recombinaison aux extrémités et donc de distinguer les zones subtélomériques, recombinant beaucoup, des télomères, qui eux ne recombinent quasiment pas. Pour autant, le séquençage reste très coûteux et ne sera pas développé dans la race Lacaune à court-terme.

Nous avions indiqué que la sélection pouvait impacter la détection des points chauds (Chan *et al.,* 2012), bien que normalement la méthode de Li et Stephen, que nous avons utilisée, n'est pas impactée. Cependant, pour être sûr de ne pas avoir d'effet de la sélection, il pourrait être intéressant d'estimer les taux de recombinaison populationnels et de rechercher les points chauds grâce à la méthode de Chan *et al.* (2012), qui corrige pour les effets de sélection. Cette méthode pourrait nous permettre d'obtenir un nombre de points chauds plus similaire à celui obtenu chez l'Homme.

Si nous pouvions créer des cartes de recombinaison historique dans d'autres races (s'il y avait suffisamment d'individus génotypés sur la puce 600K ou d'individus séquencés), nous pourrions faire une comparaison des points chauds historiques entre ces différentes races. Cela permettrait en effet de voir si les races partagent les mêmes points chauds, où s'ils sont complètement différents entre les populations, ce qui a été observé entre l'Homme et le chimpanzé (Auton et al., 2012). De telles observations pourraient amener à une discussion sur l'évolution de la recombinaison et des points chauds entre les populations. Les données de

séquence obtenues dans la population marocaine pourraient servir pour cette étude.

I. 2. b. *Utilisation de méthodes moléculaires*

Il existe différentes méthodes pour étudier directement les phénotypes de recombinaison, notamment le « sperm-typing » ou le « chip-seq », que je vous ai présentées dans le « Chapitre I ».

La méthode de « chip-seq » nécessite l'utilisation de tissus et notamment de testicules. Ces derniers sont ainsi broyés et les protéines spécifiques de la recombinaison, *SPO11* et *MLH1* sont ensuite détectées grâce à des anticorps spécifiques. Cette méthode ne permet pas d'étudier le phénotype du taux de recombinaison individuel, en revanche elle permet de connaître la localisation de la recombinaison sur le génome et ainsi de détecter des points chauds méiotiques. Cependant, c'est une méthode qui reste relativement coûteuse : quelques milliers d'euros pour l'étude d'un individu et il sera donc difficile de construire des cartes de recombinaison à l'échelle de la population avec cette méthode. Pour la mettre en application chez la Lacaune, il faudrait pouvoir récupérer des testicules de béliers, ce qui n'est pas forcément faisable, en particulier sur les béliers en reproduction dans les élevages ou dans les stations de contrôle des schémas.

La méthode de « sperm-typing », quant à elle peut donner accès à un phénotype de recombinaison individuel. Pour rappel, il s'agit d'une technique permettant, après amplification par *PCR* et génotypage de spermatozoïdes, de caractériser des points chauds et leur utilisation par l'individu dont on a récupéré les spermatozoïdes. Avec cette technique il serait possible d'évaluer la proportion de crossing-overs tombant dans les points chauds, donc leur utilisation par l'individu, correspondant au nombre d'évènements de recombinaison ayant lieu par cellule (Wang *et al.*, 2012). Cependant, pour pouvoir comparer cette utilisation des points chauds entre les individus et mettre en évidence de manière statistique une différence, il faut étudier un nombre suffisant d'animaux. Or, c'est une méthode extrêmement coûteuse, bien plus que le « chip-seq », elle est donc quasiment impossible à utiliser pour étudier l'ensemble d'une population et donc l'observation d'une possible variation des phénotypes entre les individus. Malgré le coût, elle serait quand même plus « facile » à mettre en œuvre chez la Lacaune, car il n'y a pas besoin d'avoir accès aux testicules, seulement à la semence.

Chez les ovins, bien que *PRDM9* existe, il n'a pas encore été montré comme ayant un impact sur le déterminisme des points chauds. Il serait donc intéressant de savoir si *PRDM9* détermine

également les points chauds du mouton ou si, comme pour le chien (Auton *et al.*, 2013) ou les oiseaux (Zahn, 2015), il est inactivé et les points chauds ovins seraient donc soumis à un autre déterminisme et à une autre distribution. En effet, chez les espèces présentant une absence de *PRDM9*, la recombinaison et les points chauds ont tendance à se regrouper à proximité des promoteurs des gènes, où la chromatine est plus ouverte et accessible (Campbell *et al.*, 2016). Ces points chauds semblent également être plus stables que ceux contrôlés par *PRDM9* et ne seraient donc pas soumis au « paradoxe des points chauds ». Ainsi, deux espèces d'oiseaux ne possédant pas *PRDM9* et séparées depuis des dizaines de millions d'années, présentent de nombreux points chauds similaires (Singhal *et al.*, 2015). Cependant, le mouton étant assez proche de la vache, espèce pour laquelle *PRDM9* a été montré comme impactant les points chauds et leur usage (Sandor *et al*., 2012), on aurait plutôt tendance à penser que les points chauds ovins sont également régis par *PRDM9*. Les méthodes moléculaires, bien qu'elles ne permettent d'étudier que quelques individus, pourraient cependant indiquer si l'espèce étudiée est une espèce à *PRDM9* ou pas. En effet, comme nous l'avons indiqué dans le « Chapitre I », la distribution de la recombinaison est très différente en fonction de si *PRDM9* agit ou non sur la recombinaison (Przeworski, 2016). Pour les espèces à *PRDM9,* la recombinaison a majoritairement lieu à l'écart des *TSS* et des promoteurs, alors qu'elle a lieu dans les *TSS* pour les autres. Les méthodes moléculaires, puisqu'elles permettent de connaître la distribution de la recombinaison sur le génome, pourraient ainsi indiquer si les ovins utilisent *PRDM9* ou non.

II. Préciser le déterminisme génétique

II. 1. *Le taux de recombinaison individuel*

Pour améliorer et préciser la détection du déterminisme génétique, il pourrait être intéressant d'imputer sur de la séquence. Nous avions ainsi essayé d'utiliser ces informations de séquence pour améliorer notre imputation et apporter encore plus de précision à notre *GWAS*. Cependant, les moutons marocains et les Lacaune sont deux races trop éloignées et il n'a donc pas été possible d'utiliser ces données de séquence pour préciser le déterminisme génétique du taux de recombinaison. Pourtant, si nous avions plusieurs dizaines de Lacaune intégralement séquencés, nous pourrions grandement améliorer notre imputation pour la recherche des

déterminismes génétiques. Ainsi, Kadri *et al.* (2016) ont utilisé 337 animaux intégralement séquencés, et provenant en grande majorité de la race Holstein pour préciser leurs analyses *QTLs*. Cela leur a permis d'obtenir une cartographie fine de six *QTLs* qu'ils avaient préalablement identifiés, ce qui n'avait pas été possible de réaliser dans l'étude de Sandor *et al.* (2012) où ils n'ont pu séquencer que quelques animaux choisis pour quatre *QTLs* ; ce qui est un travail lourd et coûteux.

La taille de l'échantillon à prendre en compte dans une *GWAS* dépend de plusieurs critères ; notamment l'effet du *QTL* recherché, la *MAF* des marqueurs conservés, le nombre de marqueurs utilisés, ou encore le déséquilibre de liaison entre les individus (Hong et Park, 2012). Lorsque le déséquilibre de liaison est bien connu dans la population et que le phénotype est plutôt expliqué par quelque loci, il est possible de n'utiliser que quelques centaines d'individus (Korte et Farlow, 2013). Cependant, une *GWAS* sera toujours plus précise lorsque le nombre d'individus étudié est important. La mise en place de la sélection génomique en Lacaune va permettre d'augmenter d'année en année le nombre de données disponibles. En effet, chaque année, entre 2 000 et 3 000 mâles Lacaune sont génotypés. Ceci conduira, à terme, à avoir un jeu de données en 50K très conséquent qui permettra de refaire les analyses et de préciser les *QTLs* déjà détectés, voire d'en proposer de nouveaux.

Dans le projet *Romane Ite Domum* que je vous ai présenté précédemment, seulement quelques dizaines de Romane seront phénotypés. Ce n'est donc pas suffisant pour la recherche de *QTLs*. En revanche, Il serait intéressant de pouvoir génotyper en 600K de nombreuses familles qui permettraient de réaliser de manière statistique une *GWAS* sur le phénotype étudié. Ou encore, d'utiliser des données de séquence, mais la collecte de telles données coûte très cher, ainsi pour le projet *Romane Ite Domum*, le génotypage de seulement une trentaine de familles coûte déjà près de 50 000 euros.

II. 2. *Le biais d'usage des points chauds*

Nous avons vu que nos deux jeux de données familiaux et populationnels en race Lacaune ne suffisent pas pour étudier le biais d'usage des points chauds et le déterminisme génétique de ce phénotype. De plus, c'est un phénotype difficile à étudier, car difficile à définir.

Il est possible d'avoir recours à des méthodes indirectes pour étudier le biais d'usage des points chauds. Le phénotype considéré serait alors la proportion de crossing-overs qui tombe réellement dans des points chauds (Coop *et al.*, 2008). Ils ont cherché à estimer la proportion de crossing-overs qui ont effectivement lieu dans des points chauds. Pour cela, ils ont utilisé une méthode de vraisemblance qui prend en compte la possibilité qu'un crossing-over chevauche un point chaud par hasard. Cette probabilité a été estimée en réalisant 100 distributions aléatoires du taux de recombinaison selon une loi Normale et en comptant le nombre de crossing-overs, résultant de ces simulations et qui chevauchent un point chaud.

Cette estimation prend ainsi en compte l'incertitude sur la localisation des points chauds et sur la localisation des crossing-overs puisque le chevauchement dû au hasard est à la fois influencé par la largeur des points chauds estimés et par la taille des intervalles dans lesquels sont déterminés les crossing-overs (Coop *et al.*, 2008). Lorsque ce phénotype est connu, on peut rechercher une différence d'usage des points chauds par les individus (Coop *et al.*, 2008). Des conditions particulières sont néanmoins nécessaires pour que cette méthode puisse fonctionner ; il faut notamment que la grande majorité des points chauds soit commune aux différents individus afin de pouvoir comparer leur utilisation, cela nécessite d'avoir un allèle *PRDM9* majoritaire dans la population, tout en ayant suffisamment d'allèles minoritaires différents avec des effets importants pour observer une variabilité de l'utilisation des points chauds entre les individus. Un tel dispositif est vrai chez l'Homme européen, mais pas chez le chimpanzé (Auton *et al.*, 2012).

Chez la vache, ce sont plutôt des « fenêtres chaudes » de recombinaison qui ont été étudiées (quelques dizaines de Kb) (Sandor *et al.*, 2012, Ma *et al.*, 2015). Comme nous l'avons dit dans le « Chapitre I. », il est possible de faire cette étude dans ces fenêtres, car l'augmentation de la recombinaison est corrélée à la densité en points chauds. Cependant, cela reste un phénotype très indirect qui ne correspond pas forcément au biais d'usage des points chauds.

De plus, pour étudier ce phénotype, il faut être sûr des points chauds détectés. Nous avons utilisé 51 animaux génotypés sur une puce 600K pour repérer nos points chauds, cependant, nous nous sommes aperçus que la résolution n'était pas optimale et restait de l'ordre de plusieurs kilobases, alors que les points chauds correspondent à des zones de moins de 2 Kb. Afin d'obtenir une meilleure résolution, il faudrait avoir beaucoup plus d'individus génotypés sur la puce 600K, de l'ordre de la centaine.

Pour ce qui est du déterminisme génétique, il semble que le biais d'usage des points chauds soit oligo-génique et qu'un seul gène majeur détermine ce phénotype : le gène *PRDM9*. L'effet du *QTL* serait alors très fort. Nous nous sommes ainsi demandés quel nombre d'individus il nous faudrait pour avoir suffisamment de puissance pour détecter *PRDM9* sur la puce 50K. Nous avons utilisé le programme R « luo.ld.power » du package « ldDesign » (Ball, 2003) pour répondre à cette question. Il s'agit d'une fonction prenant en entrée :

- le nombre d'individus étudiés.
- les fréquences des allèles aux marqueurs et aux *QTLs*.
- le coefficient de déséquilibre de liaison D.
- l'héritabilité au *QTL*.
- le risque α.

Nous avons utilisé l'héritabilité disponible en bovin : 0,21 (Sandor *et al.*, 2012), ce qui est possible car le déterminisme génétique du phénotype biais d'usage peut être considéré comme monogénique. Nous avons fixé les fréquences alléliques à 0,5 (les marqueurs de la 50K étant relativement communs). Nous avons calculé **D** selon la formule suivante :

$$\mathbf{D = \sqrt{p*(1-p)*q*(1-q)*r^2}} \text{ avec :}$$

- **p** : fréquence allélique au marqueur.
- **q** : fréquence allélique au *QTL*.
- $\mathbf{r^2}$: statistique permettant d'estimer le déséquilibre de liaison.

Le r^2 lié au déséquilibre de liaison est fixé à 0,2 (valeur obtenue chez les bovins, car non connue en ovin). Ainsi, pour un risque α de 0,01, nous observons qu'avec un jeu de données similaire au nôtre (350 animaux), nous avons une puissance de détection de 84%. Si on double le jeu de données, la puissance est alors de 99%. La *GWAS* permettrait ainsi d'étudier un potentiel déterminisme génétique pouvant expliquer la variation du biais d'usage, à condition d'avoir le bon phénotype.

Un nouveau projet va bientôt voir le jour, il s'agit du projet « Vargoat ». Il permettra de séquencer 225 races de chèvre françaises, ainsi que 175 races internationales et 119 races issues de l'African Goat Improvement Network qui seront séquencées aux Etats-Unis. Les objectifs principaux de ce projet seront notamment d'explorer la diversité génétique, repérer des traces de

sélection, d'adaptation et de domestication et de rechercher des gènes candidats, voire des mutations causales dans les différentes races de chèvres. Grâce à la disponibilité de ces séquences, il sera également possible de réaliser l'étude de la variation du taux de recombinaison dans cette espèce, ce qui n'a pas encore été réalisé à ce jour.

III. Intérêts de l'étude de la recombinaison en sélection

Bien que le taux de recombinaison ne soit pas un phénotype directement observable, nous avons pu voir qu'il était partiellement héritable et sous déterminisme génétique chez de nombreuses espèces de rente, telles que les ovins. Il pourrait donc être sélectionné s'il a un intérêt pour les professionnels de la filière (sélectionneurs, éleveurs).

III. 1. *Utilisation de la recombinaison pour la création de nouvelles puces*

Comme nous l'avons indiqué dans le « Chapitre IV. », il est possible d'utiliser les cartes de recombinaison génétique pour créer des puces basse densité qui peuvent être utilisées ensuite pour l'imputation. L'utilisation des distances génétiques pour la constitution de ces puces augmente la qualité de l'imputation et rendrait donc possible la création de nouvelles puces avec peu de marqueurs.

Il pourrait également être intéressant de faire tourner les logiciels d'imputation, notamment FImpute ou Impute, en leur ajoutant une carte génétique, avec les distances en cM et voir si cela peut effectivement améliorer l'imputation.

III. 2. *Utilisation du taux de recombinaison pour augmenter la réponse à la sélection*

III. 2. a. *Peut-on utiliser la recombinaison pour améliorer la réponse à la sélection ?*

La réponse à la sélection, considérée en génétique quantitative comme le changement de moyenne phénotypique entre la génération parentale et les descendants issus des individus sélectionnés, dépend de plusieurs critères : la précision de la sélection, l'intervalle entre les

générations, l'intensité de sélection et la quantité de variation génétique (Battagin *et al.*, 2016). L'utilisation des informations génomiques a permis d'augmenter la précision de la sélection, de diminuer les intervalles de générations, grâce à une évaluation précise de la ségrégation mendélienne et d'augmenter l'intensité de la sélection, ce qui permet une réduction des coûts d'évaluation d'un individu. Cependant, il est difficile d'agir sur la réponse à la sélection ; aujourd'hui, il reste peu de pistes d'améliorations possibles des coûts ou de l'intervalle de génération sans modifier les techniques de production. Une solution possible serait d'augmenter la quantité de variation génétique afin d'améliorer cette réponse à la sélection.

Cette quantité de variation génétique est impactée par la taille de la population étudiée. Chez les animaux de rente, les populations sont généralement de taille limitée et la quantité de variation génétique dépend du nombre de variants d'un caractère quantitatif (*QTV*), de leur fréquence et de la taille de leurs effets (Falconer et Mackay, 1996), mais également du degré de liaison entre les *QTVs* ; si tous les *QTVs* ségrégent indépendamment les uns des autres, et/ou s'il n'y pas de sélection, la variation génétique sera plus importante (Battagin *et al.*, 2016). Cependant, l'indépendance est rarement observée étant donné le peu d'évènements de recombinaison qui a lieu sur un chromosome.

L'idée que le taux de recombinaison augmenterait la variation de la « fitness » n'est pas nouvelle : elle a été proposée par Weismann en 1889 ; si le taux de recombinaison était sélectionné, cela permettrait d'augmenter la fitness dans la population (Weismann, 1889). De précédentes études ont permis de montrer que de hauts taux de recombinaison permettaient une meilleure réponse à la sélection (Korol et Iliadi, 1994) et que le taux de recombinaison pouvait être augmenté lorsqu'il était corrélé à un caractère directement sélectionné (Otto et Barton, 2001). Puisque nous avons montré que la recombinaison génétique des ovins était partiellement sous contrôle génétique, il serait possible d'améliorer génétiquement l'aptitude à la recombinaison, en sélectionnant notamment pour des allèles favorables (Mészáros *et al.*, 2014). La recombinaison semble en effet sélectionnable car les espèces domestiques, végétales ou animales, ont des taux de recombinaison plus élevés que les espèces sauvages (Ross-Ibarra, 2004, Ollivier, 1995). Sélectionner sur la recombinaison permettrait donc d'augmenter à chaque génération la variation génétique des caractères d'intérêt zootechniques et donc d'améliorer, à court- et long-terme, la réponse à la sélection (Battagin *et al.*, 2016).

Une étude a été réalisée par Battagin *et al.* (2016) afin de tester l'effet de la recombinaison sur la réponse à la sélection. Pour cela, ils ont réalisé des simulations permettant de tester différents taux de recombinaison sur l'ensemble du génome (de 0,1 cM/Mb à 20 cM/Mb) sur la réponse à la sélection, la perte de diversité génétique et l'intensité de la sélection. Ils ont ainsi montré que l'augmentation de la recombinaison permettait effectivement d'améliorer la réponse à la sélection et de réduire la perte de diversité génétique, de plus cela permettait de conserver une intensité de sélection assez forte, avec un même pool d'animaux.

III. 2. b. *Impacts du taux de recombinaison sur la sélection en Lacaune*

Au vu de l'étude menée par Battagin *et al.* (2016), il serait intéressant de savoir si la réponse à la sélection peut également être améliorée par l'utilisation du taux de recombinaison en Lacaune. Dans l'étude de Battagin *et al.*, (2016), 9 taux de recombinaison sont étudiés : 0,1 cM/Mb, 0,25 cM/Mb, 0,5 cM/Mb, 1 cM/Mb (considéré comme le témoin, car c'est globalement le taux de recombinaison estimé chez les animaux d'élevage), 2 cM/Mb, 5 cM/Mb, 10 cM/Mb, 15 cM/Mb et 20 cM/Mb. Comme nous l'avons montré précédemment, la Lacaune a un taux de recombinaison proche de 1,5 cM/Mb.

D'après l'étude, une augmentation du taux de recombinaison s'accompagne d'une augmentation de la réponse à la sélection et d'une réduction sur la perte des variances géniques et génétiques. Cependant, ces résultats ne sont réellement très significatifs que pour des taux de recombinaison très élevés : une augmentation de la réponse à la sélection de 28,7% et une variance génétique de 0,46 sur 40 générations pour un taux de recombinaison de 10 cM/Mb par exemple. En Lacaune, les 10% d'animaux qui recombinent le plus, ont un taux moyen de recombinaison proche de 2,5 cM/Mb. Sélectionner ces animaux permettrait d'avoir une augmentation de la réponse à la sélection proche de 15% et une variance génétique de 30% (Battagin *et al.* 2016), mais sur 40 générations. Pour 10 générations, le gain sur la réponse à la sélection est beaucoup plus faible : autour de 3%.

Le phénotype du taux de recombinaison individuel n'est pas mesurable directement, il faudrait passer soit par des génotypages soit par l'identification de mutations causales. Plusieurs gènes candidats ont été déterminés chez la Lacaune et expliquent 40% de la variance génétique, notamment *RNF212*. Il pourrait donc être intéressant d'utiliser la sélection sur les gènes repérés

comme ayant un impact sur la recombinaison, et plus particulièrement *RNF212*. Si l'on sélectionnait, en Lacaune, uniquement des individus homozygotes pour l'allèle favorisant la recombinaison au meilleur *SNP* associé à *RNF212*, on pourrait augmenter de près de 5 le nombre de crossing-overs, c'est-à-dire passer d'un taux de recombinaison moyen de 1,5 cM/Mb à 1,71 cM/Mb, soit une augmentation de seulement 14%, ce qui est cohérent avec ce qui a été observé chez la vache où la fixation de polymorphismes favorisant la recombinaison pour *RNF212* ne permettrait d'augmenter la recombinaison que de 14% (Sandor *et al.*, 2012).

Augmenter le taux de recombinaison améliore donc la réponse à la sélection et réduit également la perte des variances, mais pour des taux de recombinaison extrêmes qu'il n'est pas possible d'obtenir en Lacaune aujourd'hui. Cependant, certains animaux ont un taux de recombinaison entre 2 et 2,5 cM/Mb. Il pourrait donc être intéressant de les étudier plus en détail afin de savoir si eux-mêmes, ou leur descendance, ont développé des tares dues à ces taux de recombinaison relativement forts et, si ce n'est pas le cas, les utiliser permettrait quand même d'améliorer la réponse à la sélection de près de 15% et peut-être même de jouer sur l'intensité de sélection. Enfin, l'utilisation de la recombinaison pourrait être un nouveau moyen de maintenir la variation génétique dans une population sous sélection directe, car elle permettrait d'augmenter la variation parmi les individus sélectionnés. Cela se ferait en cassant les déséquilibres gamétiques de phases négatifs qui peuvent se créer entre *QTVs* (Battagin *et al.*, 2016). A court-terme, il y aura donc plus de variation à sélectionner puisqu'elle sera plus importante parmi les gamètes et donc parmi les candidats à sélectionner. A long-terme, une plus grande variation sera obtenue puisque la consanguinité ne sera pas détériorée au niveau des *QTVs*. En revanche, pour la sélection génomique, l'existence de forts taux de recombinaison n'a pas que des avantages. En effet, la sélection génomique s'appuie sur les corrélations entre les *SNPs* et les variants causaux et le déséquilibre de liaison. Or, l'augmentation du taux de recombinaison réduit ces corrélations et donc l'efficacité de la sélection génomique (Battagin *et al.*, 2016). Il sera donc nécessaire d'avoir des outils de génotypage plus denses pour une même précision d'index si on utilise des individus qui recombinent plus.

En revanche, avoir un très fort taux de recombinaison n'est pas forcément recherché en élevage, cela peut avoir des effets biologiques indésirables ; de trop forts taux de recombinaison peuvent augmenter les risques de mutations ou de réarrangements chromosomiques, souvent

associés avec des maladies génétiques (Inoue et Lupski, 2002). De plus, il y a des limitations d'ordre « mécanique » à l'augmentation du taux de recombinaison. Les Mammifères sont ainsi très souvent contraints pour n'avoir qu'un seul crossing-over par chromosome ou par bras chromosomique (Coop et Przeworski, 2007), sauf le mouton domestique qui peut avoir jusqu'à 1,3 crossing-over par bras (Maddox *et al.*, 2001).

III. 3. *La recombinaison accélère l'introgression génique*

L'introgression génique consiste à inclure dans le génome d'une race receveuse A, une et une seule mutation d'un gène favorable G, issu d'une autre race B, race donneuse (Hospital et Elsen, 1992). Classiquement, l'introgression génique s'effectue par un croisement entre les races A et B, suivi d'une série de rétro-croisements (« backcross ») entre les descendants portant le gène G et les individus de la race A (Hospital et Elsen, 1992). Ces « backcross » permettent de « purifier » le fond génétique des individus issus des « backcross » afin qu'il soit similaire à celui des individus de la race receveuse A. L'utilisation des marqueurs a permis de grandement améliorer l'efficacité de l'introgression génique en permettant de choisir des reproducteurs porteurs du gène G, puis en permettant de choisir des reproducteurs porteurs du gène G et proches du « type génétique » de A (Servin et Hospital, 2002). Cela a permis d'obtenir en 6 générations de « backcross » un fond génétique correspondant pour 99,2% à celui de la race receveuse.

Il serait intéressant de savoir si l'utilisation d'une recombinaison plus efficace pourrait encore aider à améliorer l'introgression génique, c'est-à-dire permettre de purifier la race receveuse plus rapidement (en utilisant moins de « backcross ») et à moindre coût.

Pour cela, il faudrait pouvoir comparer la rapidité de la purification (obtention d'un fond génétique proche de 99%) lorsqu'on utilise des animaux receveurs, issus d'une même race, mais présentant des taux de recombinaison différents : faible, intermédiaire et fort. Chez la Lacaune, les taux de recombinaison varient entre 20 et 45 crossing-overs par méiose ; trois groupes d'animaux pourraient donc être créés : taux de recombinaison faible (<20 crossing-overs par méiose), intermédiaire (entre 20 et 40 crossing-overs par méiose) et fort (>40 crossing-overs par méiose). Un même gène serait introgressé dans ces 3 groupes et on évaluerait par simulation à chaque

génération de « backcross » la proportion de fond génétique de la race receveuse obtenue. On s'attendrait ainsi à ce que le groupe d'individus présentant les plus forts taux de recombinaison atteigne 99% de fond génétique plus rapidement que les deux autres groupes, car il a été montré que de forts taux de recombinaison facilitent l'introgression en conduisant à un remplacement complet des gènes de la race donneuse, excepté le gène d'intérêt introgressé (Tanaka, 2010).

Il est quand même important de se demander si l'introgression génique a encore vraiment lieu d'être, étant donné qu'aujourd'hui, l'édition des génomes vient de voir le jour. L'édition du génome est une technique de biologie moléculaire qui modifie le génome en utilisant des enzymes particulières, les endonucléases de restriction, puis en ajoutant ou retirant des morceaux d'*ADN* spécifiques. Ces techniques permettent notamment de fixer des gènes d'intérêt dans une population. L'édition génomique a donc le même but que l'introgression génique, mais permet de l'obtenir quasiment instantanément et pas après plusieurs générations de « backcross ». La combinaison de l'utilisation de l'édition des génomes et de la recombinaison peut également aider à la détection de gènes candidats, voire de mutations causales (Sadhu *et al.*, 2016). En effet, la localisation de ces gènes et de ces variants dépend des évènements de recombinaison qui brisent les liaisons entre les marqueurs. De plus, la résolution spatiale de cette cartographie des gènes est limitée par le taux de recombinaison. Ces différents éléments impliquent que la majorité des gènes et mutations causales expliquant les phénotypes restent non identifiés (Sadhu *et al.*, 2016). Pour pallier ce problème, il est possible d'utiliser l'édition des génomes afin d'augmenter la recombinaison dans des zones d'intérêt et ainsi permettre une meilleure localisation des gènes candidats et des mutations causales, ce qui a été réalisé chez la levure (Sadhu *et al.*, 2016). L'utilisation de l'édition des génomes, si elle est un jour acceptée par la communauté, pourrait donc être un moyen pour augmenter la recombinaison, car elle permettrait de créer artificiellement des *DSBs*, et donc des crossing-overs.

III. 4. *Conclusions*

Il n'est pas possible aujourd'hui d'étudier, en Lacaune, le phénotype de biais d'usage des points chauds en raison d'un manque de résolution. Cependant, de nouveaux projets vont voir le jour, notamment le projet *Romane Ite Domum*, qui va amener beaucoup plus de données et une meilleure précision des crossing-overs. Le séquençage se développe également de plus en plus et

pourrait donc, à terme, être utilisé pour l'étude de ce phénotype. Cela pourrait notamment se faire chez la chèvre, où la recombinaison n'a pas encore été étudiée, mais où de nombreuses données de séquence vont être disponibles grâce au projet « Vargoat ».

Des données plus importantes, notamment chez les femelles Lacaune, permettraient de continuer la comparaison de la recombinaison entre les Mammifères, et notamment avec la race Soay, où il y a un effet sexe très distinct.

De plus, la recombinaison peut avoir un véritable intérêt pratique en sélection. En effet, elle est héritable, sous déterminisme génétique et des gènes candidats ont pu être mis en évidence. Il faudrait continuer de les étudier et de les identifier, afin de pouvoir proposer des mutations causales qui seraient ensuite utilisées en sélection pour augmenter la recombinaison. Cette augmentation de la recombinaison pourrait être très intéressante en sélection génomique, notamment en améliorant la réponse à la sélection et la diversité génétique.

Conclusion Générale

Conclusion générale

L'enjeu actuel est de pouvoir continuer à sélectionner les animaux, tout en essayant au maximum de préserver la variabilité génétique et d'utiliser toutes les ressources génétiques disponibles. Répondre à ces problématiques demande donc de nouveaux moyens de sélection et notamment la compréhension du mécanisme de recombinaison méiotique. Dans le cadre de ma thèse, j'ai donc utilisé des approches statistiques indirectes basées sur le déséquilibre de liaison, des approches de génétique quantitative : *GWAS* et détection de *QTLs*, ainsi que des approches de sélection génomique et d'imputation pour tester un intérêt de la recombinaison génétique en sélection.

La race Lacaune disposait de données de génotypages suffisantes pour mener à bien cette analyse. Grâce à deux jeux de données indépendants : un jeu de données familial et un jeu de données populationnel, nous avons pu étudier la distribution de la recombinaison le long du génome de la Lacaune, ainsi que le phénotype de taux de recombinaison moyen des individus et son déterminisme génétique. Cette étude a également permis de construire des cartes génétiques de haute résolution pour la Lacaune, et donc d'apporter une mise à jour à la dernière carte datant de 2007 (Maddox et Cockett, 2007), mais également pour la race ovine Soay. Cela nous a permis de constater que les deux races ont des cartes génétiques quasiment identiques, tant en terme d'intensité de recombinaison que de distribution de la recombinaison, et la combinaison des deux permet ainsi d'affiner les cartes génétiques. En revanche, même si elles ont des cartes génétiques très similaires, leur déterminisme génétique de ces phénotypes de recombinaison est différent.

Nous avons ainsi découvert que la Lacaune avait des patrons de recombinaison communs aux autres Mammifères, et nous avons pu mettre en évidence des régions particulières, notamment avec le chromosome 10 qui est le chromosome qui recombine le moins sur tout le génome et qui présente une zone similaire à celle de l'Homme, riche en *AT* avec une recombinaison quasi nulle. Il pourrait donc être intéressant de poursuivre les analyses de cette région, notamment en regardant l'état d'ouverture de la chromatine ou encore les marques épigénétiques et voir s'il peut y avoir un lien entre ces caractéristiques et cette très faible recombinaison.

Avec le jeu de données populationnel nous avons également pu détecter des points chauds

de recombinaison, ce qui n'avait pas encore été montré chez le mouton. A partir de ces points chauds, nous avons recherché des motifs pouvant être reconnus par *PRDM9*, cependant cela n'a pas été concluant avec les données actuelles, il faudra donc sûrement attendre de nouveaux projets qui permettront d'obtenir plus d'animaux génotypés avec la puce haute densité ou d'obtenir des données de séquences. Nous avons quand même pu conclure à l'existence du gène *PRDM9* chez le mouton et nous avons pu montrer qu'il était dans un scaffold non assemblé qu'il faudrait pouvoir ajouter à la fin du chromosome 1.

Puisque nous possédions deux jeux de données indépendants pour l'étude et la caractérisation de la recombinaison, il était également intéressant de les combiner. Cela a permis de montrer qu'ils étaient hautement corrélés et dans les régions où leur taux divergeaient, nous avons pu mettre en évidence des signatures de sélection, majoritairement déjà identifiées dans la littérature.

Enfin, à l'aide du jeu de données familial, nous avons pu étudier le phénotype de taux de recombinaison individuel. Comme pour la majorité des espèces étudiées, nous avons constaté que les individus présentaient une variation du taux de recombinaison qui était partiellement expliquée par un déterminisme génétique, puisque le phénotype est héritable à 23%. Deux *QTLs* principaux ont été mis en évidence suite à une *GWAS* faite après imputation : sur les chromosomes 6 et 7. Un gène candidat positionnel très intéressant a pu être découvert sur le chromosome 6 : *RNF212*. Il a déjà été associé à la variation du taux de recombinaison dans de nombreuses autres espèces et représentait notre signal le plus fort. Pour cela, nous avons décidé de rechercher de potentielles mutations causales au sein de ce gène, qui avaient des effets équivalents aux *SNPs* découverts ; des variants étaient donc très associés à notre signal, faisant de *RNF212* un sérieux candidat fonctionnel également. Nous avons également mis en évidence *HEI10* sur le chromosome 7. Il interagit avec *RNF212* et il serait intéressant d'étudier de manière plus approfondie le *QTL* du chromosome 7, car la région semble assez mal assemblée : de nombreuses séquences répétées et un trou dans l'assemblage. Il pourrait donc être intéressant d'essayer de reséquencer cette zone, notamment avec la technique PacBio, afin de mieux la préciser et donc d'affiner notre résultat de *GWAS*.

Le deuxième phénotype de recombinaison communément étudié, la variation inter-individuelle de la localisation de la recombinaison, n'a malheureusement pas pu être traité dans

cette thèse étant donné notre trop faible résolution et notre manque de données. Nous avons cependant suggéré plusieurs pistes d'améliorations possibles, notamment avec l'arrivée du projet *Romane Ite Domum*.

La prise en compte de la recombinaison génétique peut aussi avoir un intérêt en élevage et en sélection. J'ai notamment présenté une potentielle application concrète de l'utilisation des cartes génétiques pour la création de puces basse densité qui permettent d'améliorer l'efficacité l'imputation et qui conduiraient donc à la création d'outils de sélection avec moins de marqueurs qu'actuellement, et donc à faible coût pour les professionnels de l'élevage.

Les résultats obtenus dans le cadre de ma thèse sont donc un premier pas vers une meilleure compréhension de la recombinaison génétique chez le mouton. Ils ont également démontré que la recombinaison avait un intérêt pratique pour la sélection génomique. De plus, il existe donc trois espèces proches pour lesquelles des données sont ou vont être disponibles : les bovins chez qui il y a énormément de données, les ovins qui commencent à être bien caractérisés à la suite de l'étude de Johnston *et al.* (2016) et de ma thèse, et les caprins pour lesquels la sélection génomique va se mettre en place dans les deux principales races : Alpine et Saanen. Il serait donc très intéressant de pouvoir étudier l'évolution de la recombinaison entre ces différentes espèces.

Bibliographie

- Adler, ID. (1996). Comparison of the duration of spermatogenesis between male rodents and humans. *Mutation Research,* 352, pp. 169-172.
- Ahlawat S., Sharma P., Sharma R., Arora R. et De S. (2016) a. Zinc finger domain of the *PRDM9* gene on chromosome 1 exhibits high diversity in ruminants but its paralog *PRDM7* contains multiple disruptive mutations. *PLoS ONE* 11(5): e0156159. https://doi.org/10.1371/journal.pone.0156159.
- Ahlawat S., Sharma P., Sharma R., Arora R., Verma NK. *et al.* (2016) b. Evidence of positive selection and concerted evolution in the rapidly evolving *PRDM9* zinc finger domain in goats and sheep. *Animal Genetics*, 47, pp. 740-751.
- Anderson LK., Reeves A., Webb LM. et Ashley T. (1999). Distribution of crossing over on mouse synaptonemal complexes using immunofluorescent localization of *MLH1* protein. *Genetics*, 151, pp. 1569-1579.
- Argueso JL., Wanat J., Gemici Z. et Alani E. (2004). Competing crossover pathway act during meiosis in *Saccharomyces cerevisiae*. *Genetics*, 159, pp. 1259-1269.
- Arnheim N., Calabrese P. et Tiemann-Boege I. (2007). Mammalian meiotic recombination hot spots. *Annual Review of Genetics*, 41, pp. 369-399.
- Astle W. et Balding DJ. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24, pp. 451-471.
- Astruc JM., Baloche G., Larroque H., Beltran de Heredia I., Labatut J. *et al.* (2012). La sélection génomique des ovins laitiers en France : stratégies, premiers résultats des évaluations génomiques et perspectives. *Rencontres, Recherches, Ruminants*. 19, pp. 81-84.
- Auton A., Fledel-Alon A., Pfeifer S., Venn O., Segurel L. *et al.* (2012). A fine-scale chimpanzee genetic map from population sequencing. *Science*, 336, pp. 193-198.
- Auton A., Li YR., Kidd J., Oliveira K., Nadel J. *et al.* (2013). Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet.* 9(12):e1003984. doi :10.1371/journal.pgen.1003984.
- Baier B., Hunt P., Broman KW. et Hassold T. (2014). Variation in genome-wide levels of meiotic recombination is established at the onset of prophase in mammalian males. *PLoS Genet.* 10: e1004125.

- Baker SM., Plug AW., Prolla TA., Bronner CE., Harris AC. *et al.* (1996). Involvements of mouse *MLH1* in *DNA* mismatch repair and meiotic crossing over. *Nature Genetics*, 13, pp. 336-342.
- Baker CL., Kajita S., Walker M., Saxl RL., Raghupathy N. *et al.* (2015). *PRDM9* drives evolutionary erosion of hotspots in *Mus musculus* through haplotype-specific initiation of meiotic recombination. *PLoS Genet*, 11(1): e1004916.
- Balding DJ. (2006). A tutorial on statistical methods for population association studies. *Nature Review Genetics*, 7, pp. 781-791.
- Ball RD. (2003). Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. *Forest Research.*
- Battagin M., Gorjanc G., Faux AM., Johnston SE. et Hickey J. (2016). Effect of manipulating recombination rates on response to selection in livestock breeding programs. *Genetics, Selection, Evolution*, 48:44.
- Bates D., Mächler M., Bolker B. et Walker S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, pp. 1-48.
- Baudat F. et Nicolas A. (1997). Clustering of meiotic double-strand breaks on yeast chromosome III. *Proceedings of the National Academy of Sciences of the United States of America*, 94, pp. 5213-5218.
- Baudat F. et de Massy B. (2007). Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Research*, 15, pp. 565-577.
- Baudat F., Imai Y. et de Massy B. (2013). Meiotic recombination in mammals: localization and regulation. *Nature Reviews Genetics*, 14, pp. 794-806.
- Berg IL., Neumann R., Lam KWG., Sarbajna S., Odenthal-Hesse L. *et al.* (2010). *PRDM9* variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics*, 42, pp. 859-863.
- Bishop DK. et Zickler D. (2004). Early decision: meiotic crossover interference prior to stable strand exchange and synapsis. *Cell*, 117, pp. 9-15.
- Blat Y., Protacio RU., Hunter N. et Kleckner N. (2002). Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell*, 111, pp. 791-802.
- De Boer E., Jasin M. et Keeney S. (2013). Analysis of recombinants in female mouse meiosis. *Methods in Molecular Biology*, 957, pp. 19-45.

- Bolormaa S., Gore K., Van der Werf JHJ., Hayes BJ. et Daetwyler HD. (2015). Design of a low-density *SNP* chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. *Animal Genetics*, 46, pp. 544-556.
- Boulton A., Myers RS. et Redfield RJ. (1997). The hotspot conversion paradox and the evolution of meiotic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 94, pp. 8058-8063.
- Brown TA. (2002). Genomes. 2Nd edition. *Wiley-Liss*.
- Brunschwig H., Levi L., Bend-David E., Williams RW., Yakir B. *et al.* (2012). Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics*, 191, pp. 757-764.
- Buard J. et de Massy B. (2007). Playing hide and seek with mammalian meiotic crossover hotspots. *Trends in Genetics*, 23, pp. 301-309.
- Buard J. et Vergnaud G. (1994). Complex recombination events at the hypermutable minisatellite *CEB1 (D2S90)*. *EMBO Journal*, 13, pp. 3203-3210.
- Campbell CL., Bhérer C., Morrow BE., Boyko AR. et Auton A. (2016). A pedigree-based map of recombination in the domestic dog genome. *Genes, Genomes, Genetics*, X, pp.1-10.
- Capilla L., Montserrat GC. et Ruiz-Herrera, A. (2016). Mammalian meiotic recombination: a toolbox for genome evolution. *Cytogenetic and Genome Research*, 150, pp. 1-16.
- Carpenter ATC. (1979). Recombination nodules and synaptonemal complex in recombination-defective females of *Drosophila melanogaster*. *Chromosoma*, 75, pp. 259-292.
- Chagné D., Crowhurst RN., Troggio M., Davey MW., Gilmore B. *et al.* (2012). Genome-wide *SNP* detection, validation and development of an 8K *SNP* array for apple. *PLoS ONE* 7(2): e31745.
- Chan AH., Jenkins PA. et Song YS. (2012). Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genet.* 8(12): e1003090. https://doi.org/10.1371/journal.pgen.1003090.
- Cheung VG., Burdick JT., Hirschmann D., Morley M. (2007). Polymorphic variation in human meiotic recombination. *American Journal of Human Genetics*, 80, pp. 526-530.
- Chowdhury R., Bois PRJ., Feingold E., Sherman SL. et Cheung VG. (2009). Genetic analysis of variation in human meiotic recombination. *PLoS Genet.* 5(9):e1000648. doi :10.1371/journal.pgen.1000648.
- Cirulli et., Kliman RM. et Noor MAF. (2007). Fine-scale crossover rate heterogeneity in

Drosophila pseudoobscura. *Journal of Molecular Evolution*, 64, pp. 129-135.

- Cohen-Zinder M., Seroussi E., Larkin DM., Loor JJ., Everts-van der Wind A. *et al.* (2005). Identification of a missense mutation in the bovine *ABCG2* gene with a major effect on the *QTL* on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research*, 15, pp. 936-944.
- Coop G. et Myers SR. (2007). Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genet.* 3(3): e35. doi:10.1371/journal.pgen.0030035.
- Coop G. et Przeworski M. (2007). An evolutionary view of human recombination. *Nature Review of Genetics*, 36, pp. 1203-1206.
- Coop G., Wen X., Ober C., Pritchard JK. et Przeworski M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319, pp. 1395-1398.
- Cromie GA. et Smith GR. (2007). Branching out: meiotic recombination and its regulation. *Trends in Cell Biology,* 9, pp. 448-455.
- Darvasi A., Weinreb A., Minke V., Weller JI. et Soller M. (1993). Detecting marker-*QTL* linkage and estimating *QTL* gene effect and map location using a saturated genetic map. *Genetics*, 134, pp. 943-951.
- Dassonneville R., Fritz S., Ducrocq V. et Boichard D. (2012). Imputation performances of 3 low-density marker panels in beef and dairy cattle. *Journal of Dairy Science*, 95, pp. 4136-4140.
- Davies B., Hatton E., Altemose N., Hussin JG., Pratto F. *et al.* (2016). Re-engineering the zinc fingers of *PRDM9* reverses hybrid sterility in mice. *Nature*, 530, pp. 171-176.
- Druet T. et Georges M. (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics*, 184, pp. 789-798.
- Duchemin SI., Colombani C., Legarra A., Baloche G., Larroque H. *et al.* (2012). Genomic selection in the French Lacaune dairy sheep breed. *Journal of Dairy Science*, 95, pp. 2723-2733.
- Dumont BL., Broman KW. et Payseur BA. (2009). Variation in genomic recombination rates among heterogeneous stock mice. *Genetics*, 182, pp. 1345-1349.
- Edlinger B. et Schlögelhofer. (2011). Have a break: determinants of meiotic *DNA* double strand break (*DSB*) formation and processing in plants. *Journal of Experimental Botany*, 65, pp. 1545-1563.

- Falconer DS. et Mackay TFC. (1996). Introduction to quantitative genetics. 4ème édition. *Harlow : Longman Group Ltd.*
- Fan QQ., Xu F., White MA. et Petes TD. (1997). Competition between adjacent meiotic recombination hotspots in the yeast *Saccharomyces cerevisiae*. *Genetics*, 145, pp. 3661-36670.
- Fariello MI., Servin B., Tosser-Klopp G., Rupp R., Moreno C. *et al.* (2014). Selection signatures in worldwide sheep populations. *PLoS One*, 9:e103813.
- Ferrer-Admetlla A., Liang M., Korneliussen T. et Nielsen R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, 31, pp. 1275-1291.
- Fledel-Alon A., Leffler EM., Guan Y., Stephens M., Coop G. *et al.* (2011). Variation in human recombination rates and its genetic determinants. *PLoS One*, 6(6): e20321. doi:10.1371/journal.pone.0020321.
- Franklin FC., Higgins JD, Sanchez-Moran E., Armstrong SJ., Osman KE. *et al.* (2006). Control of meiotic recombination in *Arabidopsis*: role of the *MUTL* and *MUTS* homologues. *Biochemical Society Transactions*, 34, pp. 542-544.
- Galtier N., Piganeau G., Mouchiroud D. et Duret L. (2001). *GC*-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159, pp. 2907-2911.
- Game JC., Sitney KC., Cook VE. et Mortimer RK. (1989). Use of a ring chromosome and pulsed-field gels to study interhomolog recombination, double-strand DNA breaks and sister chromatid exchange in yeast. *Genetics*, 123, pp. 695-713.
- Gardiner-Garden M. et Frommer M. (1987). *CpG* islands in vertebrate genomes. *Journal of Molecular Biology*, 196, pp. 261-282.
- Georges M. (2007). Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. *Annual Review of Genomics and Human Genetics*, 8, pp. 131-162.
- Groenen MAM., Wahlberg P., Foglio M., Cheng HH., Megens HJ. *et al.* (2009). A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Research*, 19, pp. 510-519.
- Groenen MAM., Archibald AL., Uenishi H., Tuggle CK., Takeuchi Y. *et al.* (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, 491, pp. 393-398.
- Guan Y. et Stephens M. (2008). Practical issues in Imputation-Based Association Mapping.

PLoS Genet., 4(12) : e1000279.

- Haldane J. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. New College, Oxford.
- Haldane JBS. (1934). Methods for the detection of autosomal linkage in man. *Annals of Human Genetics*, 6, pp. 26-65.
- Hoffman ER. et Borts RH. (2004). Meiotic recombination intermediates and mismatch repair proteins. *Cytogenetic Genome Research*, 107, pp. 232-248.
- Holliday R. (1964). A mechanism for gene conversion in fungi. *Genetical Research*, 5, pp. 282-304.
- Hollingsworth NM. et Brill SJ. (2004). The *Mus81* solution to resolution: generating meiotic crossovers without Holliday junctions. *Genes & Development*, 18, pp. 117-125.
- Holloway JK., Booth J., Edelmann W., McGowan CH. et Cohen PE. (2008). *MUS81* generates a subset of *MLH1-MLH3*-independent crossovers in mammalian meiosis. *PLoS Genet* 4(9): e1000186. doi:10.1371/journal.pgen.1000186.
- Hong EP. et Park JW. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, 10, pp. 117-122.
- Hospital F. et Elsen JM. (1992). Introgression génique assistée par marqueurs. *INRA Productions Animales*, hs, pp. 299-302.
- Ignatieva EV., Levitsky VG., Yudin NS., Moshkin MP. et Kolchanov NA. (2014). Genetics basis of olfactory cognition : extremely high level of *DNA* sequence polymorphism in promoter regions of the human olfactory receptor genes revealed using the 1000 Genomes Project dataset. *Frontiers in Psychology*, 5, pp. 247.
- Inoue K. et Lupski JR. (2002). Molecular mechanisms for genomic disorders. *Annual Review of Genomics and Human Genetics*, 3, pp. 199-242.
- Institut de l'Elevage, CNBL. *Résultats de Contrôle Laitier – espèce ovine*. 2013. 23 p.
- International HapMap Consortium, Frazer KA., Ballinger DG., Cox DR., Hinds DA. e*t al.* (2007). A second generation human haplotype map of over 3.1 million *SNPs*. *Nature*, 449, pp. 851-861.
- Janssens, FA. (1909). La théorie de la chiasmatypie. Nouvelle interprétation des cinèses de maturation. *La Cellule*, 25, pp. 389-411.
- Jeffreys AJ., Murray J. et Neumann R. (1998). High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Molecular Cell*, 2,

pp. 267-273.

- Jeffreys AJ. et Neumann R. (2005). Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Human Molecular Genetics*, 14, pp. 2277-2287.
- Jeffreys AJ. et Neumann R. (2009). The rise and fall of a human recombination hot spot. *Nature Genetics*, 41, pp. 625-629.
- Jensen-Seaman MI., Furey TS., Payseur BA., Lu Y., Roskin KM. *et al.* (2004). Comparative recombination rates in the rat, mouse and human genomes. *Genome Research*, 14, pp. 528-538.
- Jones GH. et Franklin CH. (2006). Meiotic crossing-over : obligation and interference. *Cell*, 126, pp. 246-248.
- Johnston SE., Gratten J., Berenos C., Pilkington JG., Clutton-Brock TH. e*t al.* (2013). Life history trade-offs at a single locus maintain sexually selected genetic variation. *Nature*, 502, pp. 93-95.
- Johnston SE., Bérénos C., Slate J. et Pemberton J. (2016). Conserved genetic architecture underlying individual recombination rate variation in a wild population of Soay Sheep (*Ovis aries*). *Genetics*, 203, pp. 583-598.
- Kaback DB., Barber D., Mahon J., Lamb J. et You J. (1999). Chromosome size-dependent control of meiotic reciprocal recombination in *Saccharomyces cerevisiae* : the role of crossover interference. *Genetics*, 152, pp. 1465-1486.
- Kadri N., Harland C., Faux P., Cambisano N., Karim L. *et al.* (2016). Coding and noncoding variants in *HFM1*, *MLH3*, *MSH4*, *MSH5*, *RNF212*, and *RNF212B* affect recombination rate in cattle. *Genome Research*, 26, pp. 1323-1332.
- Kari V., Shchebet A., Neumann H. et Johnsen SA. (2011). The *H2B* ubiquitin ligase *RNF40* cooperates with *SUPT16H* to induce dynamic changes in chromatin structure during *DNA* double-strand break repair. *Cell Cycle*, 10, pp. 3495-3504.
- Kaur T. et Rockman MV. (2014). Crossover heterogeneity in the absence of hotspots in *Caenorhabditis elegans. Genetics*, 196, pp. 137-148.
- Keeney S. (2001). Mechanism and control of meiotic recombination initiation. *Current Topics in Developmental Biology*, 52, pp. 1-53.
- Kijas J., Lenstra J., Hayes B., Boitard S., Porto Neo L. *et al.* (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol* 10(2): e1001258.

- Kim ES., Elbeltagy AR., Aboul-Naga AM., Rischkowsky B., Sayre B. *et al.* (2016). Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. *Heredity*, 116, pp. 255-264.
- Klug A. (2010). The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Quarterly Reviews of Biophysics*, 43, pp. 1-21.
- Koehler KE., Cherry JP., Lynn A., Hunt PA et Hassold TJ. (2002). Genetic control of mammalian meiotic recombination I Variation in exchange frequencies among males from inbred mouse strains. *Genetics*, 162, pp. 1297-1306.
- Kong A., Gudbjartsson DF., Sainz J., Jonsdottir GM., Gudjonsson SA. *et al.* (2002). A high-resolution recombination map of the human genome. *Nature Genetics*, 31, pp. 241-247.
- Kong A., Barnard J., Gudbjartsson DF., Thorleifsson G., Jonsdottir G. *et al.* (2004). Recombination rate and reproductive success in humans. *Nature Genetics*, 36, pp. 1203-1206.
- Kong A., Thorleifsson G., Stefansson H., Masson G., Helgason A. *et al.* (2008). Sequence variants in the *RNF212* gene associate with genome-wide recombination rate. *Science*, 319, pp. 1398-1401.
- Kong A., Thorleifsson G., Frigge ML., Masson G., Gudbjartsson DF. *et al.* (2014). Common and low-frequency variants associated with genome-wide recombination rate. *Nature Genetics*, 46, pp. 11-16.
- Korol AB. et Iliadi KG. (1994). Increased recombination frequencies from directional selection for geotaxis in *Drosophila*. *Heredity*, 72, pp. 64-68.
- Korte A. et Farlow A. (2013). The advantages and limitations of trait analysis with *GWAS* : a review. *Plant Methods*, pp. 9-29.
- Lake CM. et Hawley RS. (2013). *RNF212* marks the spot. *Nature Genetics*, 45, pp. 228-229.
- Lamb NE., Sherman SL. et Hassold TJ. (2005). Effect of meiotic recombination on the production of aneuploidy gametes in humans. *Cytogenetic and Genome Research*, 111, pp. 250-255.
- Li M. et Brill SJ. (2005). Roles of *SGS1*, *MUS81*, and *RAD51* in the repair of lagging-strand replication defects in *Saccharomyces cerevisiae*. *Current Genetics*, 48, pp. 213-225.
- Li N. et Stephens M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165, pp. 2213-2233.

- Lodish H., Berk A., Zipursky SL. (2000). Section 12.5 : Recombination between Homologous DNA Sites: Double-Strand Breaks in DNA Initiate Recombination". *Molecular Cell Biology* (4th ed.). W. H. Freeman and Company.
- De Los Santos T., Hunter N., Lee C., Larkin B., Loidl J. *et al.* (2003). The *Mus81/Mms4* endonuclease acts independently of double-Holliday junction resolution to promote a distinct subset of crossovers during meiosis in budding yeast. *Genetics*, 164, pp. 81-94.
- Lynch M. et Walsh B. (1998). Genetics and analysis of quantitative traits. *Sinauer Assoc.* Sunderland, Mass.
- Lyrakou S., Mantas D., Msaouel P., Baathalah S., Shrivastav P. *et al.* (2007). Crossover analysis using immunofluorescent detection of *MLH1* foci in frozen-thawed testicular tissue. *Reproductive BioMedicine Online*, 15, pp. 99-105.
- Ma L., O'Connell JR., VanRaden PM., Shen B., Padhi A. *et al.* (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS Genet.* 11(11):e1005387. doi :10.1371/journal.pgen.1005387.
- Maddox JF., Davies KP., Crawford AM., Hulme DJ., Vaiman D. *et al.* (2001). An enhanced linkage map of the sheep genome comprising more than 1000 loci. *Genome Research*, 11, pp. 1275-1289.
- Maddox JF. et Cockett NE. (2007). An update on sheep and goat linkage maps and other genomic resources. *Small Ruminant Research*, 70, pp. 4-20.
- Mancera E., Bourgon R., Brozzi A., Huber W. et Steinmetz LM. (2008). High-resolution mapping of meiotic crossovers and non-crossover in yeast. *Nature*, 454, pp. 479-485.
- Martinez-Perez E. et Colaiacovo MP. (2009). Distribution of meiotic recombination events: talking to your neighbors, *Current Opinion in Genetics & Development*, 19, pp. 105-112.
- Mary N., Barasc H., Ferchaud S., Billon Y., Meslier F. *et al.* (2014). Meiotic recombination analyses of individual chromosomes in male domestic pigs (*Sus scrofa domestica*). *PLoS One*, 9(6): e99123. doi:10.1371/journal.pone.0099123.
- McClellan J. et King MC. (2010). Genetic heterogeneity in Human disease. *Cell*, 141, pp. 210-217.
- McVean GAT., Myers S., Hunt S., Deloukas P., Bentley DR. *et al.* (2004). The fine-scale structure of recombination rate variation in the Human genome. *Science*, 304, pp. 581-584.
- Mészáros G., Gorjanc G., Jenko J., Cleveland MA., Woolliams JA. *et al.* (2014). Selection on recombination rate to increase genetic gain. *Proceedings of the 1Oth world congress of genetics applied to livestock production*, Vancouver.

- Meunier J. et Duret L. (2004). Recombination drives the evolution of *GC*-content in the human genome. *Molecular Biology and Evolution*, 21, pp. 984-990.
- Moens PB., Kolas NK., Tarsounas M., Marcon E., Cohen PE. *et al.* (2001). The time course and chromosomal localization of recombination-related proteins at meiosis in the mouse are compatible with models that can resolve the early *DNA-DNA* interactions without reciprocal recombination. *Journal of Cell Science*, 115, pp. 1611-1622.
- Moreno-Romieux C., Tortereau F., Raoul J., Servin B. (2017). High density genotypes of French Sheep populations. Doi: 10.5281/zenodo.237116.
- Myers S., Bottolo L., Freeman C., McVean G. et Donnelly P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310, pp. 321-324.
- Myers S., Spencer CCA., Auton A., Bottolo L., Freeman C. *et al.* (2006). The distribution and causes of meiotic recombination in the human genome. *Biochemical Society Transactions*, 34, pp. 526-530.
- Myers S., Freeman C., Auton A., Donnelly P. et McVean G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, 40, pp. 1124-1129.
- Myers S., Bowden R., Tumian A., Bontrop RE., Freeman C. *et al.* (2010). Drive against hotspot motifs in Primates implicates the *PRDM9* gene in meiotic recombination. *Science*, 327, pp. 876-879.
- Nagamine Y., Pong-Wong R., Navarro P., Vitart V., Hayward C. *et al.* (2012). Localizing loci underlying complex trait variation using regional genomic relationship mapping. *PLoS One*, 7 : e46501.
- Nagel AC., Fischer P., Szawinski J., La Rosa MK. et Preiss A. (2012). *Cyclin G* is involved in meiotic recombination repair in *Drosophila melanogaster*. *Journal of Cell Science*, 125, pp. 5555-5563.
- Nassif N., Penney J., Pal S., Engels WR. et Gloor GB. (1994). Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Molecular and Cellular Biology*, 14, pp. 1613-1625.
- Neff MW., Broman KW., Mellersh CS., Ray K., Acland GM. *et al.* (1999). A second-generation genetic linkage map of the domestic dog, *Canis familiaris*. *Genetics*, 151, pp. 2803-2820.
- Neumann R. et Jeffreys AJ. (2006). Polymorphism in the activity of human crossover hotspots independent of local *DNA* sequence variation. *Human Molecular Genetics*, 15, pp.

1401-1411.

- Nishant, KT., Kumar C. et Rao MR. (2006). HUMHOT : a database of human meiotic recombination hot spots. *Nucleic Acids Research*, 34, pp. 25-28.
- Oliver PL., Goodstadt L., Bayes JJ., Birtle Z., Roach KC. e*t al.* (2009). Accelerated evolution of the *PRDM9* speciation gene across diverse metazoan taxa. *PLoS Genetics*, 25, e1000753.
- Ollivier L. (1995). Genetic differences in recombination frequency in the pigs (*Sus scrofa*). *Genome*, 38, pp. 1048-1051.
- O'Reilly PF., Birney E. et Balding DJ. (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Research*, 18, pp. 1304-1313.
- Otto SP. et Barton NH. (2001). Selection for recombination in small populations. *Evolution*, 55, pp. 1921-1931.
- Pabo CO., Peisach E. et Grant RA. (2001). Design and selection of novel *Cys2His2* zinc finger proteins. *Annual Review of Biochemistry*, 70, pp. 313-340.
- Paigen K. et Petkov P. (2010). Mammalian recombination hot spots: properties, control and evolution. *Nature Reviews Genetics*, 11, pp. 221-233.
- Pâques F. et Haber JE. (1999). Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews*, 63, pp. 349-404.
- Parvanov ED., Petkov PM. et Paigen K. (2010). *PRDM9* controls activation of mammalian recombination hotspots. *Science*, 327, pp. 835.
- Peace C., Bassil N., Main D., Ficklin S., Rosyara UR. *et al.* (2012). Development and evaluation of a genome-wide 6K *SNP* array for diploid sweet cherry and tetraploid sour cherry. *PLoS ONE* 7(12): e48305.
- Petes TD. (2001). Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics*, 2, pp. 360-369.
- Pompanon F., Benjelloun B., Leempoel K., Alberto FJ., Boyer F. *et al.* (2015). Etude de adaptation des petits ruminants marocains aux conditions environnementales par une approche de génomique du paysage. *Rencontres autour des Recherches sur les Ruminants*, 22, 125-128.
- Pratto F., Brick K., Khil P., Smagulova F., Petukhova GV. *et al.* (2014). Recombination initiation maps of individual human genomes. *Science*, 346, pp. 1256442-1, 1256442-9.

- Pritchard JK. et Przeworski M. (2001). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, 69, pp. 1-14.
- Przeworski M. (2016). Of mice, men and birds : meiotic recombination and its evolution. *The Allied Genetics Conference*, 13-17 Juillet 2016, Orlando, Floride.
- Qi LL., Echalier B., Chao S., Lazo GR., Butler GE. *et al.* (2004). A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics*, 168, pp. 701-712.
- Qiao H., Rao HBDP., Yang Y., Fong JH., Cloutier JM. *et al.* (2014). Antagonistic roles of ubiquitin ligase *HEI10* and *SUMO* ligase *RNF212* regulate meiotic recombination. *Nature Genetics*, 46, pp. 194-200.
- Rao HBDP., Qiao H., Bhatt SK., Bailey LRJ, Tran HD. *et al.* (2016). A *SUMO*-ubiquitin relay recruits proteasomes to chromosome axes to regulate meiotic recombination. *bioRxiv.*
- Reynolds A., Qiao H., Yang Y., Chen JK., Jackson N. *et al.* (2013). *RNF212* is a dosage-sensitive regulator of crossing-over during mammalian meiosis. *Nature Genetics*, 45, pp. 269-279.
- Ritz KR., Noor MAF. et Singh ND. (2017). Variation in recombination rate: adaptative or not ? *Trends in Genetics*, 33, pp. 364-374.
- Robert T., Nore A., Brun C., Maffre C., Crimi B. *et al.* (2016). The *TOPOVIB*-like protein family is required for meiotic *DNA* double-strand break formation. *Science*, 351, pp. 943-949.
- Rochus CM., Tortereau F., Plisson-Petit FI, Restoux G., Moreno-Romieux C. *et al.* (2017). High density genome scan for selection signatures in French sheep reveals allelic heterogeneity and introgression at adaptative loci. *BioRxiv, in review.*
- Rockman MV. et Kruglyak L. (2009). Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet*. 5(3). doi:10.1371/jounral.pgen.1000419.
- Romanienko PJ. et Camerini-Otero RD. (2000). The mouse *Spo11* gene is required for meiotic chromosome synapsis. *Molecular Cell*, 6, pp. 975-987.
- Rosa HJD. et Bryant MJ. (2003). Seasonality of reproduction in sheep. *Small Ruminant Research*, 48, pp. 155-171.
- Ross KJ., Fransz P. et Jones GH. (1996). A light microscopic atlas of meiosis in *Arabidopsis thaliana*. *Chromosome Research*, 4, pp. 507-516.
- Ross-Ibarra J. (2004). The evolution of recombination under domestication : a test of two

hypotheses. *The American Naturalist*, 163, pp. 105-112.

- Rosu S., Libuda DE. et Villeneuve AM. (2011). Robust crossover assurance and regulated interhomolog access maintain meiotic crossover number. *Science*, 334, pp. 1286-1289.
- Ruiz-Herrera A., Vozdova M., Fernandez J., Sebestova H., Capilla L. *et al*. (2017). Recombination correlates with synaptonemal complex length and chromatin loop size in bovids-insights into mammalian meiotic chromosomal organization. *Chromosoma*, DOI : 10.1007/s00412-016-0624-3.
- Rupp R., Mucha S., Larroque H., McEwan J. et Conington J. (2016). Genomic application in sheep and goat breeding. *Animal Frontiers*, 6, pp. 39-44.
- Sadhu MJ., Bloom JS., Day L. et Kruglyak L. (2016). *CRISPR*-directed mitotic recombination enables genetic mapping without crosses. *Science*, 352, pp. 1113-1116.
- Saintenac C. (2009). Analyse de la distribution des crossing-overs sur le chromosome 3B du blé tendre (*Triticum aestivum*) et des facteurs influençant cette distribution. *Plants genetics.* Université Blaise Pascal – Clermont-Ferrand II.
- Sallé G. et Moreno C. (2011). Apports de la génomique en élevage de petits ruminants. *Le Nouveau Praticien Vétérinaire : élevage et santé*. 17, pp. 14-18.
- Sandor C., Li W., Coppieters W., Druet T., Charlier C. *et al.* (2012). Genetic variants in *REC8*, *RNF212*, and *PRDM9* influence male recombination in cattle. *PLoS Genet.* 8(7). doi : 10.371/journal.pgen.1002854.
- Sargolzaei M., Chesnais JP. et Schenkel FS. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, 15.
- Sax K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics*, 8, pp. 552-560.
- Scheet P. et Stephens M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78, pp. 629-644.
- Serra H., Lambing C., Griffin CH., Topp SD., SA. M. *et al.* (2017). Massive crossover elevation via combination of *HEI10* and *recq4a recq4b* during *Arabidopsis* meiosis. *BioRxiv*.
- Servin B. et Hospital F. (2002). Optimal positioning of markers to control genetic background in marker-assisted backcrossing. *The Journal of Heredity*, 93, pp. 214-217.
- Servin B. et Stephens M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* 3(7): e114.

doi:10.1371/journal.pgen.0030114.

- Shifman S., Bell JT., Copley RR., Taylor MS., Williams RW. *et al.* (2006). A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol.* 4(12): e395. doi:10.1371/journal.pbio.0040395.
- Shin YH., Choi Y., Erdin SU., Yatsenko SA. *et al.* (2010). Hormad1 mutation disrupts synaptonemal complex formation, recombination, and chromosome segregation in mammalian meiosis. *PLoS Genet*. 7(2). doi: 10.1371/annotation/8aa656b6-55f7-4795-a441-cf243ea62175.
- Singer A., Perlman H., Yan YL., Walker C., Corley-Smith G. *et al.* (2002). Sex-specific recombination rates in Zebrafish (*Danio rerio*). *Genetics*, 160, pp. 2649-2657.
- Singhal S., Leffler EM., Sannareddy K., Turner I., Venn O. *et al.* (2015). Stable recombination hotspots in birds. *Science*, 350, pp. 928-932.
- Slabicki M., Theis M., Krastev DB., Samsonov S., Mundwiller E. *et al*. (2010). A genome-scale *DNA* repair *RNAi* screen identifies *SPG48* as a novel gene associated with hereditary spastic paraplegia. *PLoS Biol.*, 8: e1000408.
- Smeds L., Mugal CF., Qvarnström A. et Ellegren H. (2016). High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genet.*, 24: e1006044.
- Smith KN. et Nicolas A. (1998). Recombination at work for meiosis. *Current Opinion in Genetics & Development*, 8, pp. 200-211.
- Soller M., Brody T. et Genizi A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoritical and Applied Genetics*, 47, pp. 35-39.
- Stathopoulos S., Bishop JM. et O'Ryan C. (2014). Genetic signatures for enhanced olfaction in the African Mole-Rats. *PLoS ONE* 9(4): e93336.
- Steinmetz M., Minard K., Horvath S., McNicholas J., Srelinger J. *et al.* (1982). A molecular map of the immune response region from the major histocompatibility complex of the mouse. *Nature*, 300, pp. 35-42.
- Stevison LS. et Noor MAF. (2010). Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*. *Journal of Molecular Evolution*, 71, pp. 332-345.
- Stewart MN. et Dawson DS. (2008). Changing partners: moving from non-homologous to homologous centromere pairing in meiosis. *Trends in Genetics*, 21, pp. 564-573.

- Storey JD. et Tibshirani R. (2013). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100, pp. 9440-9445.
- Sturtevant AH. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology*, 14, pp. 43-59.
- Takasuga A. (2015). *PLAG1* and *NCAPG-LCORL* in livestock. *Animal Science Journal*, 87, pp. 159-167.
- Tanaka Y. (2010). Recombination and epistasis facilitate introgressive hybridization across reproductively isolated populations : a gamete-based simulation. *Evolutionary Ecology Research*, 12, pp. 1-22.
- Tapanainen JS., Aittomäki K., Min J., Vaskivuo T. et Huhtaniemi IT. (1997). Men homozygous for an inactivating mutation of the follicle-stimulation hormone (*FSH)* receptor gene present variable suppression of spermatogenesis and fertility. *Nature Genetics*, 15, pp. 205-206.
- Tortereau F., Servin B., Frantz L., Megens HJ., Milan D. *et al.* (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC-content. *BMC Genomics*, 13.
- Tortereau F., Palhière I., Moreno C., Tosser-Klopp G., Barbotte L. *et al.* (2014). Assignation de parentés pour les populations françaises de petits ruminants en sélection. *Rencontres autour des Recherches sur les Ruminants*, 21, pp. 257-260.
- Visscher PM., Brown MA., McCarthy MI. et Yang J. (2012). Five years of *GWAS* discovery. *The American Journal of Human Genetics*, 90, pp. 7-24.
- Voight BF., Kudaravalli S., Wen X. et Pritchard JK. (2006). A map of recent positive selection in the human genome. *PLoS Biol.*, 4:e72.
- Vrielynck N., Chambon A., Vezon D., Pereira L., Chelysheva L. *et al.* (2016). A *DNA* topoisomerase VI-like complex initiates meiotic recombination. *Science*, 351, pp. 939-943.
- Walker M., Billings T., Baker CL., Powers N., Tian H. *et al.* (2015). Affinity-seq detects genome-wide *PRDM9* binding sites and reveals the impact of prior chromatin modifications on mammalian recombination hotspot usage. *Epigenetics & Chromatin*, 8:31.
- Wang TF., Kleckner N. et Hunter N. (1999). Functional specificity of *MutL* homolog in yeast: evidence for three *Mlh1*-based heterocomplexes with distinct roles during meiosis in recombination and mismatch correction. *Proceedings of the National Academy of Sciences of the United States of America*, 96, pp. 13914-13919.
- Wang J., Fan HC., Behr B. et Quake SR. (2012). Genome-wide single-cell analysis of

recombination activity and de novo mutation rates in human sperm. *Cell*, 150, pp. 402-412.

- Ward JO., Reinholdt LG., Motley WW., Niswander LM., Deacon DC. *et al.* (2007). Mutation in mouse *Hei10,* an E3 Ubiquitin Ligase, disrupts meiotic crossing over. PLoS Genet 3(8): e139. doi:10.1371/journal.pgen.0030139.
- Weismann A. (1889). The significance of sexual reproduction in the theory of natural selection. *Essays upon heredity*, 2ème édition, pp. 257-342.
- Wong AK., Ruhe AL., Dumont BL., Robertson DKR., Guerrero G. *et al.* (2010). A comprehensive linkage map of the dog genome. *Genetics*, 184, pp. 595-605.
- Zenger KR., McKenzie LM. et Cooper DW. (2002). The first comprehensive genetic linkage map of a marsupial : the Tammar Wallaby (*Macropus eugeni*). *Genetics*, 162, pp. 321-330.
- Zhang Z. et Druet T. (2010). Marker imputation with low-density marker panels in Dutch Holstein cattle. *Journal of Dairy Science*, 93, pp. 5487-5494.
- Zhang Z., Ersoz E., Lai CQ., Todhunter RJ., Tiwari HK. *et al.* (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42, pp. 355-360.
- Zhou X. et Stephens M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44, pp. 821-824.
- Zhou X., Carbonetto P. et Stephens M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.*, 9: e1003264.

Annexes

Annexe 1

Variation in recombination rate and its genetic determinism in sheep (*Ovis Aries)* populations from combining multiple genome-wide datasets.

Morgane Petit[*], Jean-Michel Astruc[†], Julien Sarry[*], Laurence Drouilhet[*], Stéphane Fabre[*], Carole Moreno[*], Bertrand Servin[*]

*INRA, Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France

[†]Institut de l'Elevage, F-31326 Castanet-Tolosan, France

Short running title: Genetics of recombination in sheep populations.

Keywords: recombination rate; genetic maps; QTLs; evolution; sheep

Corresponding author:

Bertrand Servin

GenPhySE, INRA UR631, F-31326 Castanet-Tolosan, France

phone: +33 5 61 28 51 17

Email: bertrand.servin@inra.fr

Abstract

Recombination is a complex biological process that results from a cascade of multiple events during meiosis. Understanding the genetic determinism of recombination can help to understand if and how these events are interacting. To tackle this question, we studied the patterns of recombination in sheep, using multiple approaches and datasets. We constructed male recombination maps in a dairy breed from the south of France (the Lacaune breed) at a fine scale by combining meiotic recombination rates from a large pedigree genotyped with a 50K SNP array and historical recombination rates from a sample of unrelated individuals genotyped with a 600K SNP array. This analysis revealed recombination patterns in sheep similar to other mammals but also genome regions that have likely been affected by directional and diversifying selection. We estimated the average recombination rate of Lacaune sheep at 1.5 cM/Mb, identified about 50,000 crossover hotspots on the genome and found a high correlation between historical and meiotic recombination rate estimates. A genome-wide association study revealed two major loci affecting inter-individual variation in recombination rate in Lacaune, including the *RNF212* and *HEI10* genes and possibly 2 other loci of smaller effects including the *KCNJ15* and *FSHR* genes. Finally, we compared our results to those obtained previously in a distantly related population of domestic sheep, the Soay. This comparison revealed that Soay and Lacaune males have a very similar distribution of recombination along the genome and that the two datasets can be combined to create more precise male meiotic recombination maps in sheep. Despite their similar recombination maps, we show that Soay and Lacaune males exhibit different heritabilities and QTL effects for inter-individual variation in genome-wide recombination rates.

Introduction

Meiotic recombination is a fundamental biological process that brings a major contribution to the genetic diversity and the evolution of eukaryotic genomes (Baudat and al., 2013). During meiosis, recombination enables chromosomal alignment resulting in proper disjunction and segregation of chromosomes, avoiding deleterious outcomes such as aneuploidy (Hassold, Hall, and Hunt 2007).

Over generations, recombination contributes to shaping genetic diversity in a population by creating new allelic combinations and preventing the accumulation of deleterious mutations. Over large evolutionary timescales, divergence in recombination landscapes can lead to speciation: the action of a key actor in the recombination process in many mammals, the gene *PRDM9*, has been shown to have a major contribution to the infertility between two mouse species, making it the only known speciation gene in mammals today (Mihola et al. 2009).

Genetics studies on recombination were first used to infer the organisation of genes along the genome (Sturtevant 1913). With the advance in molecular techniques, more detailed physical maps and eventually whole genome assemblies are now available in many species. The establishment of highly resolutive recombination maps remains of fundamental importance for the validation of the physical ordering of markers, obtained from sequencing experiments (Groenen et al. 2012; Jiang et al. 2014). From an evolutionary perspective the relevant distance between loci is the genetic distance and recombination maps are essential tools for the genetic studies of a species, for estimation of past demography (H. Li and Durbin 2011; Boitard et al. 2016), detection of selection signatures (Sabeti et al. 2002; Voight et al. 2006), QTL mapping (Cox et al. 2009) and imputation of genotypes (Howie, Donnelly, and Marchini 2009) for genome-wide association studies (GWAS) or genomic selection. Precise recombination maps can be estimated using different approaches. Meiotic recombination rates can be estimated from the observation of markers' segregation in families. Although this is a widespread approach, its resolution is limited by the number of meioses that can be collected within a population and the number of markers that can be genotyped. Consequently highly resolutive meiotic maps have been produced in situations where large segregating families can be studied and genotyped densely (Shifman et al. 2006; Mancera et al. 2008; Groenen et al. 2009; Rockman and Kruglyak 2009; Augustine Kong et al. 2010) or by focusing on specific genomic regions (Cirulli, Kliman, and Noor 2007; Stevison and Noor 2010; Kaur and Rockman 2014). In livestock species, the recent availability of dense genotyping assays has fostered the production of highly resolutive recombination maps (Tortereau et al. 2012; Susan E. Johnston et al. 2016; S. E. Johnston et al. 2017) in particular by exploiting reference population data from genomic selection programs (Sandor et al. 2012a; Ma et al. 2015; Kadri et al. 2016a) .

Another approach to study the distribution of recombination on a genome is to exploit patterns of correlation between allele frequencies in a population (*i.e.* Linkage Disequilibrium, LD) to infer past (historical) recombination rates (McVean, Awadalla, and Fearnhead 2002; N. Li and Stephens 2003; Chan, Jenkins, and Song 2012). Because the LD-based approach exploits in essence meioses

accumulated over many generations, it can provide more precise estimates of local variation in recombination rate. For example, until recently (Pratto et al. 2014; Lange et al. 2016) this was the only indirect known approach allowing to detect fine scale patterns of recombination genome-wide in species with large genomes. Several highly recombining intervals (recombination hotspots) were detected from historical recombination rate maps and confirmed or completed those discovered by sperm-typing experiments (Simon Myers et al. 2005; Crawford et al. 2004). One important caveat of LD-based approaches is that their recombination rate estimates are affected by other evolutionary processes, especially selection that affects LD patterns unevenly across the genome. Hence differences in historical recombination between distant genomic regions have to be interpreted with caution. Despite this, historical and meiotic recombination rates usually exhibit substantial positive correlation (Rockman and Kruglyak 2009; Chan, Jenkins, and Song 2012; Brunschwig et al. 2012; J. Wang et al. 2012).

The LD-based approach does not allow to study individual phenotypes and therefore to identify directly loci influencing inter-individual variation in recombination rates. In contrast, family-based studies in human (A. Kong et al. 2008a; Chowdhury et al. 2009), Drosophila (Stevison and Noor 2010; Chan, Jenkins, and Song 2012) mice (Shifman et al. 2006; Brunschwig et al. 2012) cattle (Sandor et al. 2012a; Ma et al. 2015; Kadri et al. 2016a) and sheep (Susan E. Johnston et al. 2016) have demonstrated that recombination exhibits inter-individual variation and that this variation is partly determined by genetic factors. Two recombination phenotypes have been described: the number of crossovers per meiosis (Genome-wide Recombination Rate , GRR herein) and the fine scale localization of crossovers (Individual Hotspot Usage, IHU herein). GRR has been shown to be influenced by several genes. For example, a recent genome-wide association study found evidence for association with 6 genome regions in cattle (Kadri et al. 2016a). Among them, one of the genomic regions consistently found associated to GRR in mammals is an interval containing the *RNF212* gene. In contrast to GRR, the IHU phenotype seems mostly governed by a single gene in most mammals, *PRDM9*. This zinc-finger protein has a key role in recruiting *SPO11*, thereby directing DNA double-strand breaks (DSBs) that initiate meiotic recombination. Because *PRDM9* recognizes a specific DNA motif, the crossover events happen in hotspots carrying this motif. This *PRDM9* associated process is however not universal, as it is only active in some mammals; canids for example do not carry a functional copy of *PRDM9* and exhibit different patterns for the localization of recombination hotspots (Auton et al. 2013).

As mentioned above, recombination was studied recently in sheep (Susan E. Johnston et al. 2016), which lead to the production of precise genome-wide recombination maps, revealed a similar genetic architecture of recombination rates in sheep as in other mammals and identified two major loci affecting individual variation. Quite interestingly, one the QTL identified in this study, localized near the *RNF212* gene, was clearly demonstrated to have a sex specific effect. This study was performed in a feral population of sheep which is quite distantly related to continental populations (Kijas et al. 2012) and has not managed by humans for a long time. To understand how recombination patterns and genetic determinism can vary across populations, we conducted in this work a study in another sheep population, the Lacaune, from south of France. The Lacaune breed is the main dairy sheep population in France, its milk being mainly used for the production of Roquefort cheese. Starting in 2011, a large genotyping effort started in the breed to implement a genomic selection program (Baloche et al. 2014), and young selection candidates are now routinely genotyped for a medium density genotyping array (about 50K SNP). This constitutes a large dataset of genotyped families that can be used to study recombination, although limited to one sex as only males were used for genomic selection in this population. This dataset offers an opportunity to study variation in recombination and its genetic determinism between very diverged populations of the same species. Hence, a first objective of this study was to elucidate whether these two sheep populations had similar distribution of recombination on the genome and whether they shared the same genetic architecture of the trait, and in particular the same QTLs effects.

The second objective of this study was to compare different approaches to study recombination from independent data in the same population. To this end, in addition to the pedigree data, we exploit a sample of 51 unrelated individuals genotyped with a high-density genotyping array (about 500K SNP). While, the family data was used to establish meiotic recombination maps, the sample of densely genotyped individuals was used to create historical recombination maps of higher resolution. This offered the opportunity to evaluate to which extent sheep ancestral recombination patterns match contemporary ones.

Materials and Methods

Study Population and Genotype Data

In this work, we exploited two different datasets of sheep from the Lacaune breed: a pedigree dataset of 8,085 related animals genotyped with the medium density Illumina Ovine Beadchip® including 54,241 SNPs, and a diversity dataset of 70 unrelated Lacaune individuals selected as to represent population genetic diversity, genotyped with the high density Illumina Ovine Infinium® HD SNP Beadchip including (Rochus et al. 2017; Moreno-Romieux et al. 2017)

Standard data cleaning procedures were carried out on the pedigree dataset using plink 1.9 (Chang et al. 2015), excluding animals with call rates below 95% and SNPs with call freq below 98%. After quality controls we exploited genotypes at 46,813 SNPs and 5,940 meioses. For these animals, we only selected the sires which had their own sire known and at least 2 offspring and the sires which did not have their own sire known, but at least 4 offspring. Eventually, 345 male parents, called focal individuals (FIDs) hereafter, met these criteria: 210 FIDs had their father genotype known while the remaining 135 did not (Figure 1).

Recombination Maps

Meiotic recombination maps from pedigree data

Detection of crossovers

Crossover locations were detected using LINKPHASE (Druet and Georges 2015). From the LINKPHASE outputs (*recombination_hmm* files), we extracted crossovers boundaries. We then identified crossovers occurring in the same meiosis less than 3 Mb apart from each other (that we call double crossovers) and considered them as dubious. This number was chosen as it corresponded to clear outliers in the distribution of inter-crossover distances. They are also quite unlikely under crossover interference. We applied the following procedure: given a pair of double crossovers, we set the genotype of the corresponding offspring as missing in the region spanned by

the most extreme boundaries and re-run the LINKPHASE analysis. After this quality control step, we used the final set of crossovers identified by LINKPHASE to estimate recombination rates. This dataset consisted of 213,615 crossovers in 5,940 meioses.

Estimation of recombination rates

Based on the inferred crossover locations, meiotic recombination rates were estimated in windows of one megabase and between marker intervals of the medium SNP array using the following statistical model, inspired by (Cheung et al. 2007). For small genetic intervals such as considered here, the recombination rate (termed c in the following), is usually expressed in centiMorgans per megabase and the probability that a crossover occurs in one meiosis in an interval *j* (measured in Morgans) is $0.01c_jl_j$ where l_j is the length of the interval expressed in megabases. When considering M meioses, the expected number of crossovers in the interval is $0.01c_jl_jM$. When combining observations in multiple individuals, we want to account for the fact that they have different average numbers of crossovers per meiosis (termed R_s for individual *s*). To do so we multiply the expected number of crossovers in the interval by an individual specific factor equal to (R_s/R) where R is the average number of crossovers per meiosis among all individuals. Finally, for individual *s* in interval *j* the expected number of crossovers is $0.01c_jl_jM_sR_s/R$. Given this expected number, a natural distribution to model the number of crossovers observed in an interval is the Poisson distribution so that the number y_{sj} of crossovers observed in the interval *j* for an individual *s* is modelled as:

$$y_{sj} \mid c_j \sim \text{Poisson}(\, 0.01\, l_j\, c_j\, M_s\, R_s/R) \quad (1)$$

To combine crossovers across individuals, the likelihood for c_j is the product of poisson likelihoods from equation (1).

We then specify a prior distribution for c_j:

$$c_j \sim \Gamma(\alpha\, ,\, \beta) \quad (2)$$

To set α and ß, first raw c_j estimates are computed using the method of (Sandor et al. 2012a) across the genome and then a gamma distribution is fitted to the resulting genome-wide distribution

(Figure S1). Combining the prior (2) with the likelihoods in equation (1), the posterior distribution for c_j is:

$$c_j | \, y_{\cdot j} \sim \Gamma(\alpha + \sum_s y_{sj} \, , \beta + \sum_s 0.01 \, l_j M_s R_s / R) \quad (3)$$

As the localization of crossovers was usually not good enough to assign them with certainty to a single genomic interval, final estimates of c_j are obtained as follows:

(i) for each crossover overlapping interval j and localized within a window of size L, let x_c be an indicator variable that takes value 1 if the crossover occurred in interval j and 0 otherwise. Assuming that, locally, recombination rate is proportional to physical distance, set $P(x_c = 1) = min(l_j/L, 1)$.

(ii) Using the probability in step (i), sample x_c for each crossover overlapping interval j and set $y_{sj} = \sum_c x_c$

(iii) Given y_{sj} , sample c_j from equation (3)

For each interval considered, perform step (ii) and (iii) above 1000 times to draw samples from the posterior distribution of c_j thereby accounting for uncertainty in the localization of crossovers.

Historical recombination maps from the diversity data

The diversity data contains 70 Lacaune individuals genotyped for a High Density SNP array comprising 527,823 autosomal markers (Rochus et al. 2017). Nineteen of these individuals are FIDs in the pedigree data. To perform the LD-based analysis on individuals unrelated to the pedigree study, these individuals were therefore removed from the dataset and the subsequent analyses performed on the 51 remaining individuals. Population-scaled recombination rates were estimated using PHASE (N. Li and Stephens 2003). For computational reasons and to allow for varying effective population size along the genome, estimations were carried out in 2 Mb windows, with an additional 100 Kb on each side overlapping with neighbouring windows, to avoid border-effect in the PHASE inference. PHASE was run on each window with default options, except that the number of main iterations was increased to obtain larger posterior samples for recombination rate

estimation (option -X10) as recommended in the documentation.

From the PHASE output, 1000 samples were obtained from the posterior distribution of:

- The background recombination rate: $\rho_w = 4N_w c_w$, where N_w is the effective population size in the window, c_w is the recombination rate comparable to the family-based estimate.
- An interval specific recombination intensity λ_j, for each marker interval *j* of length l_j in the window, such that the population scaled genetic length of an interval is: $\delta_j = \rho_w \lambda_j l_j$

The medians were used as point estimates of parameters λ_j and δ_j, computed over the posterior distributions $\{\lambda_j^{(k)}, \lambda_j^{(k)} \rho_w^{(k)} l_j ;\ k \in [1,1000]\}$.

Identification of intervals harbouring crossover hotspots

Intervals that showed an outlying λ_j value compared to the genome-wide distribution of λ_j were considered as harbouring a crossover hotspot. Specifically, a mixture of Gaussian distribution was fitted to the genome-wide distribution of $log_{10}(\lambda_j)$ using the mclust R package (Fraley and Raftery 2002), considering that the major component of the mixture modelled the background distribution of λ_j in non-hotspots intervals. From this background distribution, a p-value was computed for each interval that corresponded to the null hypothesis that it does not harbour a hotspot. Finally, hotspot harbouring intervals were defined as those for which FDR(λ_j) < 5%, estimating FDR with the (Storey and Tibshirani 2003) method, implemented in the R qvalue package. This procedure is illustrated in Figure S2.

Combination of meiotic and historical recombination rates and construction of a high resolution recombination map

To construct a meiotic recombination map of the HD SNP array requires that the historical recombination rate estimates be scaled by 4 times the effective population size. Due to evolutionary pressures, the effective population size varies along the genome, so it must be estimated locally. This can be done by exploiting the meiotic recombination rate inference obtained from the pedigree data analysis as explained below.

Consider a window of one megabase on the genome, using the approach described above, we can sample values c_{jk} (window *j*, sample *k*) from the posterior distribution of the meiotic recombination

rate c_j. Similarly, using output from PHASE we can extract samples ρ_{jk} from the posterior distribution of the historical recombination rates ($\rho_j = \rho_w \lambda_j$). Now, considering that $\rho_j = 4Ne_j c_j$ where Ne_j is the local effective population size of window j, we get $log(\rho_j) = log(4Ne_j) + log(c_j)$. This justifies using a model on both c_{jk} and ρ_{jk} values:

$$y_{ijk} = \mu + x_{ijk}\,\alpha + \beta_j + \nu_{ij} + e_{ijk} \quad (4)$$

where y_{ijk} is $log(c_{jk})$ when i=1 (meiotic-recombination rate sample) and y_{ijk} is $log(\rho_{jk})$ when i=2 (historical recombination rate sample). In this model, μ estimates the log of the genome-wide recombination rate, x_{ijk}=1 if i=2 and 0 otherwise so that α estimates log(4Ne), where Ne is the average effective population size of the Lacaune population, $\mu + \beta_j$ estimates log(c_j) combining population and meiotic recombination rates, and $\alpha+(\nu_{2j}-\nu_{1j})$ estimates log($4Ne_j$). μ and α_i were considered as fixed effects while β_j and ν_{ij} were considered as independent random effects. Using this approach allows to combine in a single model LD- and pedigree-based inferences, while accounting for their respective uncertainties as we exploit posterior distribution samples.

Model (4) was fitted on 20 samples of the posterior distributions of c_j and ρ_j for all windows of one megabase covering the genome, with an additional fixed effect for each chromosome, using the lme4 R package (Bates et al. 2015). Windows lying less than 4 Mb from each chromosome end were not used because inference on c_j was possibly biased in these regions (see Results). After estimating this model, historical recombination rate estimates of HD intervals were scaled within each window by dividing them by their estimated local effective population size (*i.e.* $exp(\hat{\alpha}_j + \hat{\nu}_{2j} - \hat{\nu}_{1j})$ for window *j*). For windows lying within 4 Mb of the chromosome ends, historical recombination rate estimates were scaled using the genome-wide average effective population size exp($\hat{\alpha}_i$). This led eventually to estimates of the meiotic recombination rates, expressed in centiMorgans per megabase, for all intervals of the HD SNP array, which we termed a high resolution recombination map.

Effect of recombination hotspots on the recombination rate

For each interval of the medium density SNP array, we computed the number of significant hotspots detected as explained above and the hotspot density (number of hotspots per unit of physical distance). After having corrected for the chromosome effect, the GC content effect and for windows farther than 4 Mb of the chromosome end, we fitted a linear regression model to estimate the effect of hotspots density on the meiotic recombination rate.

Comparison with Soay sheep recombination maps and integration of the two datasets to produce new male recombination maps in Sheep

In order to compare the recombination maps in Lacaune with the previously established maps in Soay sheep (Susan E. Johnston et al. 2016), we downloaded the raw data from the dryad data repository (doi: 10.5061/dryad.pf4b7) and the additional information available on https://github.com/susjoh/GENETICS_2015_185553. As the approach used in (Susan E. Johnston et al. 2016) to establish recombination maps differs from the one used here, we chose to apply the method of this study to the Soay data to perform a comparison that would not be affected by difference in methods. As the Lacaune data consist only of male meioses, we also only considered male meioses in the Soay data. The final Soay dataset used consisted of 3,445 individuals among which were 299 male FIDs, defined as in the Lacaune analysis. After detecting crossovers with LINKPHASE, one FID exhibited a very high average number of crossover per meiosis (> 100) and was not considered in the analyses (Soay individual ID : RE4844), leaving 298 FIDs. The final dataset consisted of 88,683 crossovers in 2,609 male meiosis and was used to estimated meiotic recombination maps using the exact same approach as described above, both on intervals of one megabase and on the same intervals as the ones considered in the Lacaune meiotic maps on the medium density SNP array. Note that the Soay sheep are not necessarily polymorphic for the same markers as the Lacaune, but that our method is flexible and can nonetheless estimate recombination rates in intervals bordered by monomorphic markers: in such a case adjacent intervals will have the same estimated recombination rate. As the two populations were found to have very similar meiotic recombination maps (see Results), the two sets of crossovers were finally merged to create a combined dataset of 302,298 crossovers in 8,549 male meioses and to estimate new male sheep

recombination maps, again on one megabase intervals and on intervals of the medium density SNP array.

Genome-Wide Association Study on Recombination Phenotypes

Genome-wide Recombination Rate (GRR)

The set of crossovers detected was used to estimate the genome-wide recombination rate (GRR) of each FID in the family dataset from their observed number of crossovers per meiosis, adjusting for covariates: year of birth of the parent, considered as a cofactor with 14 levels for years spanning from 1997 to 2010 and insemination month of the offspring's ewe, treated as a cofactor with 7 levels for months spanning from February to August. We used a mixed-model for estimating the population average GRR μ, covariates fixed effects β and individual breeding values u_s, while controlling for non genetic individual specific effects a_s:

$$y_{so} = \mu + \boldsymbol{x_{so}\beta} + a_s + u_s + e_{so}$$

with $u_s \sim N(0, \boldsymbol{A\sigma}_s{}^2)$, $a_s \sim N(0, I\sigma_a{}^2)$ and $e_{so} \sim N(0, \boldsymbol{I\sigma_e{}^2})$, where y_{so} is the number of crossovers in the meiosis between FID *s* and offspring *o*, A is the pedigree-based relationship matrix between FIDs and $\boldsymbol{x_{so}}$ the line of the corresponding design matrix for observation y_{so}. We fitted this model using BLUPf90 (Misztal et al., 2002) and extracted: (i) estimates of variance components $\sigma_{\boldsymbol{e}}{}^{\boldsymbol{2}}$, $\sigma_a{}^2$ and $\sigma_s{}^2$, which allows to estimate the heritability of the trait (calculated as $\sigma_s{}^2/(\sigma_s{}^2 + \sigma_a{}^2 + \sigma_{\boldsymbol{e}}{}^{\boldsymbol{2}})$) and (ii) prediction $\widetilde{u_s}$ of GRR deviation for each FID.

Genotype Imputation

Nineteen of the 345 FIDs are present in the diversity dataset of HD genotypes. For the 336 remaining FIDs, their HD genotypes at 507,784 SNPs were imputed with BimBam (Guan and Stephens 2008; Servin and Stephens 2007) using the 70 unrelated Lacaune individuals as a panel.

To impute, BimBam uses the fastPHASE model (Scheet and Stephens 2006), which relies on methods using cluster of haplotypes to estimate missing genotypes and reconstruct haplotypes from unphased SNPs of unrelated animal. BimBam was run with 10 expectation-maximization (EM) starts, each EM was run 20 steps on panel data alone, and an additional 1 step on cohort data, with a number of clusters of 15. After imputation BimBam estimates for each SNP in each individual an average number of alleles, termed *mean genotype*, computed from the posterior distribution of the three possible genotypes. This mean genotype has been shown to be efficient for performing association tests (Guan and Stephens 2008). In subsequent analyses, we used the mean genotypes provided by BimBam of the 345 FIDs at all markers of the HD SNP array. To assess the quality of genotype imputation at the most associated regions, 10 markers of the HD SNP array, 1 in chromosome 6 associated region and 9 in the chromosome 7 associated region (see Results) were genotyped for 266 FIDs for which DNA samples were still available. We evaluated the quality of imputation for the most significant SNPs by comparing for each possible genotype its posterior probability estimated by Bimbam to the error rate implied by calling it. We observed a very good agreement between the two measures (Figure S3), which denoted good calibration of the imputed genotypes at top GWAS hits.

Single- and multi-QTLs GWAS on GRR

We first tested association of individual breeding values $\tilde{u}_s$ with mean genotypes at 503,784 single SNPs imputed with BimBam. We tested these associations using the univariate mixed-model approach implemented in the Genome-wide Efficient Mixed Model Association (Gemma) software (Zhou and Stephens, 2012). To accounts for polygenic effects on the trait, the centered genomic relationship matrix calculated from the mean genotypes was used. The p-values reported in the results correspond to the Wald test.

To go beyond single SNP association tests, we also estimated a Bayesian sparse linear mixed-model (Zhou, Carbonetto, and Stephens 2013) as implemented in Gemma. This method allows to consider multiple QTLs in the model, together with polygenic effects at all SNPs. The principle of the method is to have for each SNP *l* an indicator variable γ_l that takes value 1 if the SNP is a QTL and 0 otherwise. The strength of evidence that a SNP is a QTL is measured by the posterior probability $P(\gamma_l=1)$, called posterior inclusion probability (PIP). Note that all SNPs are included in the model when doing so. Inference of the model parameters is performed using an iterative MCMC

algorithm: the number of iterations was set to 10 millions and inference was made on samples extracted every 100 iterations. When a genome region harbors a QTL, multiple SNP in the region can have elevated PIPs. To summarize the strength of evidence for a *region* to carry a QTL, we calculated a rolling sum of PIPs over 50 consecutive SNPs using the rollsum function of the R zoo package (Zeileis and Grothendieck 2005). Given that the average physical distance between SNPs on the high-density SNP array is about 5 kilobases, this procedure interrogates the probability of the presence of a QTL in overlapping windows of approximately 250 kilobases.

For the univariate analysis, the False Discovery Rate was estimated using the ash package (Stephens 2017) and SNPs corresponding to an FDR < 10% were deemed significant and annotated. For the multivariate analysis, regions where the rolling sum of PIPs exceeded 0.15 were further annotated. The annotation of the QTL regions consisted in extracting all genes from the Ensembl annotation v87 along with their Gene Ontology (GO) annotations and interrogated for their possible involvement in recombination.

Variant Discovery and Additional Genotyping in *RNF212*

Identification and assignation of the RNF212 sheep genome sequence

The *RNF212* gene was not annotated on the *Ovis aries* v3.1 reference genome. Nevertheless, a full sequence of *RNF212* was found in the scaffold01089 of *Ovis orientalis* (assembly Oori1, NCBI accession NW_011943327). By BLAST alignment of this scaffold, ovine *RNF212* could be located with confidence on chromosome 6 in the interval OAR6:116426000-116448000 of Oari3.1 reference genome (Figure S4). This location was confirmed by BLAST alignment with the bovine *RNF212* gene sequence. We also discovered that the Oari3.1 unplaced scaffold005259 (NCBI accession JH922970) contained the central part of *RNF212* (exons 4-9) and it could be placed within a large assembly gap. Moreover, we also observed that automatically annotated non-coding RNA in the *RNF212* interval matched exonic sequence of *RNF212* (Figure S4).

Variant discovery in RNF212 in the Lacaune population

Based on the genomic sequence and structure of the *RNF212* gene annotated in *Ovis orientalis* (NCBI accession NW_011943327), a large set of primers were designed using PRIMER3 software (Table S1) for amplification of each annotated exon and some intron part corresponding to exonic region annotated in *Capra hircus* (Chir_v1.0). PCR amplification (GoTaq, Promega) with each primer pair was realized on 50ng of genomic DNA from 4 selected homozygous Lacaune animals exhibiting the GG and AA (non imputed) genotypes at the most significant SNP of the medium density SNP array of the chromosome 6 QTL (rs418933055, p-value 2.56e-17). Each PCR product was sequenced via the BigDye Terminator v3.1 Cycle Sequencing kit and analyzed on an ABI3730 sequencing machine (Applied Biosystems). Sequenced reads were aligned against the *Ovis orientalis RNF212* gene using CLC Main Workbench Version 7.6.4 (Qiagen Aarhus) in order to identify polymorphisms.

Genotyping of mutations in RNF212

The genotyping of 266 genomic DNA from Lacaune animals for the four identified polymorphisms within the ovine *RNF212* gene was done by Restriction Fragment Length Polymorphism (RFLP) after PCR amplification using dedicated primers (Table S1) (GoTaq, Promega), restriction enzyme digestion (BsrBI for SNP_14431_AG; RsaI for SNP_18411_GA; and Bsu36I for both SNP_22570_CG and SNP_22594_AG; New England Biolabs) and resolution on 2% agarose gel.

Results

High-Resolution Recombination Maps

Meiotic recombination maps: genome-wide recombination patterns

We studied meiotic recombination using a pedigree of 6,230 individuals, genotyped for a medium density SNP array (50K) comprising around 54,000 markers. After quality controls we exploited genotypes at 46,813 SNPs and identified 213,615 crossovers in 5,940 meioses divided among 345 male parents (FIDs) (see Methods). The pedigree information available varied among focal individuals (Figure 1): 210 FIDs had their father genotype known while the remaining 135 did not. Having a missing parent genotype did not affect the detection of crossovers as the average number of crossovers per meiosis in the two groups was similar (36.1 with known father genotype and 35.8 otherwise) and the statistical effect of the number of offspring on the average number of crossovers per meiosis was not significant ($p>0.23$). This can be explained by the fact that individuals that lacked father genotype information typically had a large number of offspring (17.4 on average, ranging from 4 to 111), allowing to infer correctly their haplotype phase from their offspring genotypes only. Overall, given that the physical genome size covered by the medium density SNP array is 2.45 gigabases, we estimate that the mean recombination rate in our population is about 1.5 cM/megabase.

Based on the crossovers identified, we developed a statistical model to estimate meiotic recombination rates (see methods) and constructed meiotic recombination maps at two different scales: for windows of one megabase and for each interval of the medium density SNP array. As this statistical approach allowed to evaluate the uncertainty in recombination rate estimates, we provide respectively in File S1 and S2, along with the recombination rate estimates in each interval, their posterior variance and 90% credible intervals. Graphical representation of the meiotic recombination maps of all autosomes are given in File S3.

The recombination rate on a particular chromosome region was found to depend highly on its position relative to the telomere and to the centromere for metacentric chromosomes, *i.e.* chromosomes 1, 2 and 3 in sheep (Figure S5). Specifically, for acrocentric and metacentric

chromosomes, recombination rate estimates were elevated near telomeres and centromeres, but very low within centromeres. In our analysis, recombination rate estimates were found low in intervals lying within 4 megabases of chromosome ends. While this could represent genuine reduction in recombination rates near chromosome ends it is also likely due to crossovers being undetected in our analysis as only few markers are informative to detect crossovers at chromosome ends. In the following analyses, we therefore did not consider regions lying within 4 Mb of the chromosomes ends.

From local recombination rate estimates in 1 Mb windows or medium SNP array intervals, we estimated chromosome specific recombination rates (Figure S6). Difference in recombination rates between chromosomes was relatively well explained by their physical size, larger chromosomes exhibiting smaller recombination rates. Even after accounting for their sizes, some chromosomes showed particularly low (chromosomes 9, 10 and 20) or particularly high (chromosomes 11 and 14) recombination rates. In low recombining chromosomes, large regions had very low recombination, between 9 and 14 Mb on chromosome 9, 36 and 46 Mb on chromosome 10 and between 27 and 31 Mb on chromosome 20. In highly recombining chromosomes, recombination rates were globally higher on chromosome 14, while chromosome 11 exhibited two very high recombination windows between 7 Mb and 8 Mb and between 53 and 54 Mb. In addition, we found, consistent with the literature, that GC content was quite significantly positively correlated with recombination rate both in medium SNP array intervals (p-value < 10^{-16}, r=0.20) and in 1 Mb intervals (p-value < 10^{-16} , r=0.28).

Estimation of historical recombination rates and identification of crossover hotspots

We used a different dataset, with 51 unrelated individuals from the same Lacaune population genotyped for the Illumina HD SNP array (600K) comprising 527,823 autosomal SNPs after quality controls. Using a multipoint model for LD patterns (N. Li and Stephens 2003), we estimated, for each marker interval of the HD SNP array, historical recombination rates (see Methods). Compared to meiotic maps, these estimates offer a greater precision as they in essence exploit meioses cumulated over many generations. However, the historical recombination rates obtained are scaled by the effective population size ($\rho = 4\, N_e\, c$ where Ne is the effective population size

and c the meiotic recombination rate) which is unknown, and may vary along the genome due to evolutionary pressures, especially selection. Thanks to the higher precision in estimation of recombination rate, LD-based recombination maps offer the opportunity to detect genome intervals likely to harbour crossover hotspots. A statistical analysis of historical recombination rates (see Methods) identified about 50,000 intervals exhibiting elevated recombination intensities (Figure S2) as recombination hotspots, corresponding to an FDR of 5%. From our historical recombination map, we could conclude that 80% crossover events occurred in 40% of the genome and that 60% of crossover events occurred in only 20% of the genome (Figure S7).

High-resolution recombination maps combining family and population data

Having constructed recombination maps with two independent approaches and having datasets in the same population of Lacaune sheep allowed first to evaluate to which extent historical crossover hotspots explain meiotic recombination, and second to estimate the impact of evolutionary pressures on the historical recombination landscape of the Lacaune population. We present our results on these questions in turn.

We studied whether variation in meiotic recombination can be attributed to the historical crossover hotspots detected from LD patterns only. For each interval between two adjacent SNPs of the medium density array, we (i) extracted the number of significant historical hotspots and (ii) calculated the historical hotspot density (in number of hotspots per unit of physical distance). We found both covariates to be highly associated with meiotic recombination rate estimated on family data (r=0.15 with hotspot density ($p < 10^{-16}$) and r=0.19 with the number of hotspots ($p<10^{-16}$)). These correlations hold after correcting for chromosome and GC content effects (respectively r=0.14 ($p<10^{-16}$) and r=0.18 ($p<10^{-16}$)). Figure 2 illustrates this finding in two one-megabase intervals from chromosome 24, one that exhibits a very high recombination rate (7.08 cM/Mb) and the second a low one (0.46 cM/Mb). In this comparison, the highly recombining window carries 36 recombination hotspots while the low recombinant one exhibits none. As the historical background recombination rates in the two windows are similar (0.7/Kb for the one with a high recombination rate, and 0.2/Kb for the other), the difference in recombination rate between these two regions is largely due to their contrasted number of historical crossover hotspots.

In order to study more precisely the relationship between historical and meiotic recombination rates, we fitted a linear mixed model (see Methods) that allowed to estimate the average effective population size of the population, the correlation between meiotic and historical recombination rates and to identify genome regions where historical and meiotic recombination rates were significantly different. We found the effective population size of the Lacaune population to be about 7,000 individuals and a correlation of 0.73 between meiotic and historical recombination rates (Figure 3). We discovered 7 regions where historical recombination rates were much lower than meiotic ones and 3 regions where they were much higher (Table 1, Figure S8). Seven of these 10 regions have extreme recombination rates compared to other genomic regions. To quantify to which extent a window is extreme, we indicate in Table 1, for each window, the proportion of the genome with a lower recombination rate (q_w). For 6 of these 7 regions, the historical recombination rate is more extreme than the meiotic rate: four regions have very low meiotic recombination rate and even lower historical recombination rates (the two regions on chromosome 3 and two regions on chromosome 10, between 36-37 megabases and between 42-44 megabases); two regions have very high meiotic recombinations rates and even higher historical recombination rates (on chromosome 12 and on chromosome 23). For these six regions, the discrepancy between meiotic recombination and historical recombination estimates can be explained by the fact that we used a genome-wide prior in our model to estimate meiotic recombination rates that has the effect of shrinking our estimates toward the mean. Because historical estimates were not shrunk in the same way, for these six outlying regions the two estimates did not concur and it is possible that our meiotic recombination rate estimates were slightly over (resp. under) estimated.

Out of the four remaining outlying windows, three had a low historical recombination rate but did not have particularly extreme meiotic recombination rates, so that the effect of shrinkage is not likely to explain the discrepancy between meiotic and historical recombination rates. Indeed, these three regions corresponded to previously identified selection signatures in sheep: a region on chromosome 6 spanning 2 intervals between 36 and 38 megabases contains the *ABCG2* gene, associated to milk production (Cohen-Zinder et al. 2005), and the *LCORL* gene associated to stature (recently reviewed in (Takasuga 2015)). This region has been shown to have been selected in the Lacaune breed (Fariello et al. 2014; Rochus et al. 2017); a region spanning one interval on chromosome 10, between 29 and 30 megabases contains the *RXFP2* gene, associated to polledness and horn phenotypes (Susan E. Johnston et al. 2013) and found to be under selection in many sheep breeds (Fariello et al. 2014); and a region on chromosome 13 between 63 and 64 megabases that

contains the *ASIP* gene responsible for coat color phenotypes in many breeds of sheep (Norris and Whan 2008), again previously demonstrated to have been under selection. For these three regions, we explain the low historical recombination estimates by a local reduction of the effective population size due to selection.

Finally, one of the three regions with a high historical recombination rate, on chromosome 20 between 28 and 29 megabases had a low meiotic recombination rate, so that the effect of shrinkage cannot explain the discrepancy. This region harbours a cluster of olfactory receptors genes and its high historical recombination rate could be explained by selective pressure for increased genetic diversity in these genes (*i.e.* diversifying selection), a phenomenon which has been shown in other species (*e.g.* pig (Groenen et al. 2012), human (Ignatieva et al. 2014), rodents (Stathopoulos, Bishop, and O'Ryan 2014)). Finally, we used the meiotic recombination rates to scale the historical recombination rate estimates and produce high-resolution recombination maps on the HD SNP array (Supporting File S4).

Improved male recombination maps by combining Lacaune and Soay sheep data

Recently, recombination maps have been estimated in another sheep population, the Soay (Susan E. Johnston et al. 2016). Soay sheep is a feral population of ancestral domestic sheep living on an island located northwest of Scotland. The Lacaune and Soay populations are genetically very distant, their genome-wide Fst, calculated using the sheephapmap data (Kijas et al. 2012), being about 0.4. Combining our results with results from the Soay offered a rare opportunity to study the evolution of recombination over a relatively short time scale as the two populations can be considered separated at most dating back to domestication, about 10,000 years ago. The methods used in the Soay study are different from those used here, but the two datasets are similar, although the Soay data has fewer male meioses (2,604 vs. 5,940 in the present study). In order to perform a comparison that would not be affected by differences in estimation methods, we ran the method developed for the Lacaune data to estimate recombination maps on the Soay data. As the Soay study showed a clear effect of sex on recombination rates, we estimated recombination maps on male meioses only. Figure 4 presents the comparison of recombination rates between the two populations in marker intervals of the medium density SNP array. The left panel shows that the two populations exhibit very similar recombination rates ($r = 0.82$, $p<10^{-16}$), although Soay

recombination rates appear higher for low recombining intervals (c < 1.5 cM/Mb in gray on the figure). We explain this by the shrinkage effect of the prior, that is more pronounced in the Soay as the dataset is smaller: the right panel on Figure 4 shows that the posterior variance of the recombination rates are clearly higher in Soays for low recombining intervals while they are similar for more recombining intervals. Overall, our results on the comparison of the recombination maps in the two populations are consistent with the two populations having the same amplitude and distribution of recombination on the genome. We therefore analyzed the two populations together to create new male recombination maps based on 302,298 crossovers detected in 8,549 meioses (Supporting file S5). Combining the two dataset together lead to a clear reduction in the posterior variance of the recombination rates, i.e. an increase in their precision (Figure S9).

Genetic Determinism of Genome-wide Recombination Rate in Lacaune sheep

Our dataset provides information on the number of crossovers for a set of 5,940 meioses among 345 male individuals. Therefore, it allows to study the number of crossovers per meiosis (GRR) as a recombination phenotype.

Genetic and environmental effects on GRR

We used a linear mixed-model to study the genetic determinism of GRR. The contribution of additive genetic effects was estimated by including a random FID effect with covariance structure proportional to the matrix of kinship coefficients calculated from pedigree records (see Methods). We also included environmental fixed effects in the model: year of birth of the FID and insemination month of the ewe for each meiosis. We did not find significant differences between the FID year of birth, however the insemination month of the ewe was significant ($p = 1.3\ 10^{-3}$). There was a trend in increased recombination rates from February to May followed by a decrease until July and a regain in August although the number of inseminations in August is quite low, leading to a high standard error for this month (Figure S10). Based on the estimated variance components (Table 2), we estimated the heritability of GRR in the Lacaune male population at 0.23.

Genome-wide association study identifies three major loci affecting GRR in Lacaune sheep

The additive genetic values of FIDs, predicted from the above model were used as phenotypes in a genome-wide association study. Among the 345 FIDs with at least two offsprings, the distribution of the phenotype was found to be approximately normally distributed (Figure S11). To test for association of this phenotype with SNPs markers, we used a mixed-model approach correcting for relatedness effects with a genomic relationship matrix (see Methods). Using our panel of 70 unrelated Lacaune, we imputed the 345 FIDs for markers of the HD SNP array. With these imputed genotypes, we performed two analyses. The first was an association test with univariate linear mixed models, which tested the effect of each SNP in turn on the phenotype (results in Supporting File S6) the second fitted a Bayesian sparse linear mixed model, allowing multiple QTLs to be included in the model (results in Supporting File S7).

Figure 4 illustrates the GWAS results: the top plot shows the p-values of the single SNP analysis and the bottom plot, the posterior probability that a region harbours a QTL, calculated on overlapping windows of 20 SNPs. The single SNP analysis revealed six significant regions (FDR < 10 %): two on chromosome 1, one on chromosome 6, one on chromosome 7, one on chromosome 11 and one chromosome 19. Regions of chromosome 6 and 7 exhibited very low p-values whereas the other three showed less intense association signals. The multi-QTLs Bayesian analysis was conclusive for two regions (regions on chromosome 6 and chromosome 7) while the rightmost region on chromosome 1 was suggestive (Table 3). Two additional suggestive regions were discovered on chromosome 3. Using the multi-QTL approach of (Zhou, Carbonetto, and Stephens 2013) allowed to estimate that together, QTLs explain about 40% of the additive genetic variance for GRR, with a 95% credible interval ranging from 28 to 53 %.

The most significant region was located on the distal end of chromosome 6 and corresponded to a locus frequently associated to variation in recombination rate. In our study the significant region contained 10 genes: *CTBP1*, *IDUA*, *DGKQ*, *GAK*, *CPLX1*, *UVSSA*, *MFSD7*, *PDE6B*, *PIGG* and *RNF212*. For each of these genes, except *RNF212* which was not annotated on the genome (see below), we extracted their gene ontology of the Ensembl v87 database, but none was clearly annotated as potentially involved in recombination. However, two genes were already reported as having a statistical association with recombination rate: *CPLX1* and *GAK (Augustine Kong et al.*

2014). *CPLX1* has no known function that can be linked to recombination (Augustine Kong et al. 2014) but *GAK* has been shown to form a complex with the cyclin-G, which could impact recombination (Nagel et al. 2012). However, *RNF212* can be deem a more likely candidate due to its function and given that this gene was associated with recombination rate variation in human (Chowdhury et al. 2009),(A. Kong et al. 2008b), in bovine (Sandor et al. 2012b; Kadri et al. 2016b; Ma et al. 2015) and in mouse (Reynolds et al. 2013). *RNF212* is not annotated in the sheep genome assembly oviAri3, however this chromosome 6 region corresponds to the bovine region that contains *RNF212* (Figure S4). We found an unassigned scaffold (scaffold01089, NCBI accession NW_011943327) of *Ovis orientalis musimon* (assembly Oori1) that contained the full *RNF212* sequence and that could be placed confidently in the QTL region. To confirm *RNF212* as a valid positional candidate, we studied further the association of its polymorphisms with GRR in results presented below.

The second most significant region was located between 22.5 and 23.1 megabases on chromosome 7. All significant SNPs in the region were imputed, *i.e.* the association would not have been found based on association of the medium density array alone. It matched an association signal on GRR in Soay sheep (Susan E. Johnston et al. 2016). Consistent with our finding, in the Soay sheep study, this association was only found using regional heritability mapping and not using single SNP associations with the medium density SNP array. This locus could match previous findings in cattle (association on chromosome 10 at about 20 Mb on assembly btau3.1), however the candidate genes mentioned in this species (*REC8* and *RNF212B*) were located respectively 2 and 1.5 megabases away from our strongest association signal. In addition, none of the SNPs located around these two candidate genes in cattle were significant in our analysis. Eleven genes were present in the region: *OR10G2*, *OR10G3*, *TRAV5*, *TRAV4*, *SALL2*, *METTL3*, *TOX4*, *RAB2B*, *CHD8*, *SUPT16H* and *RPGRIP1*. The study of their gene ontology, extracted from the Ensembl v87 database, revealed that none of them were associated with recombination, although *SUPT16H* could be involved in mitotic DSB repair (Kari et al. 2011). However another functional candidate, *CCNB1IP1*, also named *HEI10*, was located between positions 23,946,971 and 23,951,850 bp, about 500 Kb from our association peak. This gene is a good functional candidate as it has been shown to interact with *RNF212*: *HEI10* allows to eliminate the *RNF212* protein from early recombination sites and to recruit other recombination intermediates involved in crossover maturation (Qiao et al. 2014; Rao et al. 2016). Again SNPs located at the immediate proximity of *HEI10* did not exhibit significant

associations with GRR. Hence, our association signal did not allow to pinpoint any clear positional candidate among these functional candidates (see Figure S12). However, it was difficult to rule them out completely for three reasons. First, with only 345 individuals, our study may not be powerful enough to localize QTLs with the required precision. Second, the presence of causal regulatory variants, even at distances of several hundred kilobases is possible. Finally, the associated region of *HEI10* exhibited apparent rearrangements with the human genome, possibly due to assembly problems in oviAri3. These assembly problems could be linked to the presence of genomic sequences coding for the T-cell receptor alpha chain. This genome region is in fact rich in repeated sequences making its assembly challenging. Overall, identifying a single positional and functional candidate gene in this gene-rich misassembled genomic region was not possible based on our data alone.

Our third associated locus was located on chromosome 1 between 268,600 and 268,700 kilobases. In cattle, the homologous region, located at the distal end of cattle chromosome 1, has also been shown to be associated with GRR (Kadri et al. 2016a; Ma et al. 2015). In these studies the *PRDM9* gene has been proposed as a potential candidate gene, especially because it is a strong functional candidate given its proven effect on recombination phenotypes. In sheep, *PRDM9* is located at the extreme end of chromosome 1, around 275 megabases, 7 megabases away from our association signal (Ahlawat et al. 2016). Hence, *PRDM9* was not a good positional candidate for association with GRR in our sheep population. However, the associated region on chromosome 1 contains a single gene, *KCNJ15*, which has been associated with DNA double-strand breaks repair in human cells (Słabicki et al. 2010).

Finally, the two regions on the chromosome 3 were analyzed. The first was located between 75,162 and 75,319 kilobases and contains only one annotated gene coding for the receptor for follicle-stimulating hormone (*FSHR*). Though it does not affect recombination directly it is necessary for the initiation and maintenance of normal spermatogenesis in males (Tapanainen et al. 1997). The second region on the chromosome 3 was located between 201,198 and 201,341 kilobases but does contain any annotated gene.

Mutations in the RNF212 gene are strongly associated to Genome-wide Recombination Rate variation in Lacaune sheep

The QTL with the largest effect in our association study corresponded to a locus associated to GRR variation in other species and harbouring the *RNF212* gene. As it was a clear positional and functional candidate gene, we carried out further experiments to interrogate specifically polymorphisms within this gene. As stated above, we used the sequence information available for the *RNF212* gene from *Ovis orientalis* which revealed that *RNF212* spanned 23,7 Kb on the genome and may be composed of 12 exons by homology with bovine *RNF212*. However, mRNA annotation indicated multiple alternative exons. Surprisingly, the genomic structure of ovine *RNF212* was not well conserved with goat, human and mouse syntenic *RNF212* genes (Figure S4). As a first approach, we designed primers for PCR amplification (see Methods) and sequencing of all annotated exons and some intronic regions corresponding to exonic sequences of *Capra hircus RNF212*. By sequencing *RNF212* from 4 carefully chosen Lacaune animals homozygous GG or AA at the most significant SNP of the medium density SNP array on chromosome 6 QTL (rs418933055, p-value 2.56 10^{-17}), we evidenced 4 polymorphisms within the ovine *RNF212* gene (2 SNPs in intron 9, and 2 SNPs in exon 10). The 4 mutations were genotyped in 266 individuals of our association study. We then tested their association with GRR using the same approach as explained above (results in Supporting File S8) and computed their linkage disequilibrium (genotypic r^2) with the most associated SNPs of the high-density genotyping array (see Figure S13) (Table 4). Two of these mutations were found highly associated with GRR, their p-values being of the same order of magnitude ($p<10^{-16}$) as the most associated SNP (rs412583165), one of them was even more significant than the most significant imputed SNP ($p = 6.25\ 10^{-17}$ vs $p = 9.8\ 10^{-17}$). We found a clear agreement between the amount of LD between a mutation and the most associated SNPs and their association p-value (see Figure S13). Overall, these results showed that polymorphisms within the *RNF212* gene were strongly associated with GRR, and likely tagged the same causal mutation as the most associated SNP. This confirmed that *RNF212*, a very strong functional candidate, was also a very strong positional candidate gene underlying our association signal.

The genetic determinism of recombination differ between Soay and Lacaune males

GWAS in the Soay identified two major QTLs for GRR, with apparent sex-specific effects. These two QTLs were located in the same genomic regions as our QTLs on chromosome 6 and chromosome 7. The chromosome 6 QTL was only found significant in Soay females, while we detect a very strong signal in Lacaune males. Although the QTL was located in the same genomic region, the most significant SNPs were different in the two GWAS (Figure 6). Two possible explanations could be offered for these results: either the two populations have the same QTL segregating and the different GWAS hits correspond to different LD patterns between SNPs and QTLs in the two populations, or the two populations have different causal mutations in the same region. Denser genotyping data, for example by genotyping the RNF212 mutations identified in this work in the Soay population, would be needed to have a clear answer. For the chromosome 7 QTL, the signal was only found using regional heritability mapping (Nagamine et al. 2012) in the Soay, and after genotype imputation in our study, which makes it even more difficult to discriminate between a shared causal mutation or different causal mutations at the same location in the two populations.

Discussion

In this study, we studied the distribution of recombination along the sheep genome and its relationship to historical recombination rates. We showed that contemporary patterns of recombination are highly correlated to the presence of historical hotspots. We showed that the recombination patterns along the genome are conserved between distantly related sheep populations but that their genetic determinism of genome-wide recombination rates differ. In particular, we showed that polymorphisms within the *RNF212* gene are strongly associated to male recombination in Lacaune whereas this genomic region shows no association in Soay males. Hence, combining three datasets, two pedigree datasets in distantly related domestic sheep populations and a densely genotyped sample of unrelated animals, revealed that recombination rate and its genetic determinism can evolve at short time scales, as we discuss below.

Fine-scale Recombination Maps

In this work, we were able to construct fine-scale genetic maps of the sheep autosomes by combining two independent inferences on recombination rate. Our study on meiotic recombination from a large pedigree dataset revealed that sheep recombination exhibits general patterns similar to other mammals (Shifman et al. 2006; Chowdhury et al. 2009; Tortereau et al. 2012). First, sheep recombination rates were elevated at the chromosome ends, both on acrocentric and metacentric chromosomes. In the latter, our analysis revealed a clear reduction in recombination near centromeres. Second, recombination rate depended on the chromosome physical size, consistent with an obligate crossover per meiosis irrespective of the chromosome size. These patterns were consistent with those established in a very different sheep population, the Soay (Susan E. Johnston et al. 2016), and indeed when re-analysing the Soay data with the same approach as used in this study, the results showed a striking similarity between recombination rates in the two populations. Hence, our results show that recombination patterns were conserved over many generations and despite the very different evolutionary histories of the two populations and clear differences in the genetic determinism of GRR in males of the two populations. This similarity allowed to combine the two datasets to create more precise male sheep recombination maps than any of the two studies taken independently.

Our historical recombination maps revealed patterns of recombination at the kilobase scale, with small highly recombining intervals interspaced by more wide, low recombining regions. This result was consistent with the presence of recombination hotspots in the highly recombinant intervals. A consequence was that, as observed in other species, the majority of recombination took place in a small portion of the genome: we estimated that 80% of recombination takes place in 40% of the genome. (Kaur and Rockman 2014) suggested to use a Gini coefficient as a measure of the heterogeneity in the distribution of recombination along the genome to facilitate inter-species comparisons. When calculated on the historical recombination data, the Lacaune sheep has a coefficient of 0.52, which is similar to what is observed in Drosophila but lower than that measured in humans or mice. However, the coefficient calculated here is likely an underestimate due to our limited resolution (a few kilobases on the HD SNP array) compared to the typical hotspot width (a few hundred base-pairs). Overall, we identified 50,000 hotspot intervals which was twice the estimated number of hotspots in humans (International HapMap Consortium et al. 2007). This difference can be explained by different non mutually exclusive reasons. First, it is possible that what we detect as crossover hotspots are due to genome assembly errors and we indeed found a significant albeit moderate effect (OR 1.4) of the presence of assembly gaps in an interval on its probability of being called a hotspot. Second, our method to call hotspots could be too liberal. Indeed, a more stringent threshold (FDR=0.1%) would lead to about 25,000 hotspots, which would be similar to what is found in humans. Third, selection has been shown to impact hotspots discoveries although not with the methods we used here (Chan, Jenkins, and Song 2012). Finally, there exists the possibility that historically sheep exhibits more recombination hotspots than humans. In any case, the strong association between meiotic recombination rate and density in historical hotspots showed that our historical recombination maps were generally accurate. We tried to find enrichment in sequence motifs in the detected hotspots or specify their position relative to TSS (data not shown), but with no success mainly due to (i) the relative large hotspots intervals (about 5Kb) compared to typical hotspot motifs and (ii) the quality of the sheep genome assembly which still contains many small gaps that make such analyses difficult. Ultimately these questions would need an improved genome assembly and better resolution of crossover hotspots which should be addressed in the future from LD-based studies on resequencing data.

We combined, using a formal statistical approach, meiotic- and LD-based recombination rate estimates. Using an approach conceptually similar to that of (O'Reilly, Birney, and Balding 2008) led us to assess the impact of selection events on the sheep genome, in particular suggesting the

possibility of an effect of diversifying selection at olfactory receptors genes. Based on this comparison, the correlation between historical and meiotic recombination rates was found to be high (r = 0.7), but less than could be expected from previous results in humans, where the correlation was 97% on 5 Mb (S. Myers et al. 2006). However, it was closer to that of worms, mice or Drosophila, 69%, 47% and 50% respectively (Rockman and Kruglyak 2009, Brunschwig *et al.* 2012, Chan *et al.* 2012). Again, more precise estimates of both meiotic and historical recombination rates could change this number but other causes can be put forward.

A first explanation could come from the fact that the model we used to estimate historical recombination rates is based on the assumption of a constant effective population size, both in the past and along the genome. To allow for varying population size along the genome, we estimated the model in 2Mb intervals but there is still the possibility that varying population size in the past affect our historical recombination rate estimates, as the method has been shown to be somewhat influenced by demography although the identification of crossover hotspots much less so (N. Li and Stephens 2003). Also, as already mentioned above, selection has been shown to have substantial impact on estimation of recombination rates with other approaches (Chan, Jenkins, and Song 2012) although it has not been evaluated for the Li and Stephens model to our knowledge.

Second, our meiotic recombination maps are based on male meioses only while historical recombination rates are averaged over both male and female meioses. The fact that male and female recombination differ substantially, in particular in sheep (Susan E. Johnston et al. 2016), could also explain this relatively lower correlation.

Third, it is also possible that selective pressure due to domestication and later artificial breeding had the impact of modifying extensively LD patterns on the sheep genome, degrading the correlation between the two approaches. Indeed, the historical recombination estimates summarize ancestral recombinations that took place in the past and it is possible that recombination hotspots that were present in an ancestral sheep population are not longer active in today's Lacaune individuals. This could arise, for example, if domestication led to a reduction in the diversity of hotspots defining genes, such as *PRDM9*, and hence a reduction in the number of motifs underlying hotspots which would in turn change the distribution of recombination on the genome. This has been shown for example in Humans were patterns of recombination differ between populations due to their different diversity at *PRDM9* (Berg et al. 2010, 2011; Baudat et al. 2010). Eventually, such a phenomenon would degrade the correlation between present day recombination (measured by the meiotic recombination rates) and past recombination (measured by historical recombination rates).

Further studies on the determinism of hotspots in the sheep, its related genetic factors and their diversity would be needed to elucidate this question.

Despite these different effects, the substantial correlation between meiotic and historical recombination rates motivates the creation of scaled recombination maps that can be useful for interpreting statistical analysis of genomic data. As an illustration of the importance of fine-scale recombination maps for genetic studies, we found an interesting example in a recent study on adaptation of sheep and goat (Kim et al. 2016). In this study, a common signal of selection was found using the iHS statistic (Voight et al. 2006) in these two species (Figure 5 in (Kim et al. 2016)). This signature matches precisely the low recombining regions we identified on chromosome 10. However, the iHS statistic has been shown to be strongly influenced by variation in recombination rates, and in particular to tend to detect low recombining regions as selection signatures (Ferrer-Admetlla et al. 2014; O'Reilly, Birney, and Balding 2008). Precise genetic maps such as the one we provide in this work could thus help in annotating and interpreting such selection signals.

Determinism of Recombination Rate in sheep populations

As mentioned in the introduction two phenotypes have been studies with respect to the recombination process, but only one was studied here, Genome-wide recombination rate (GRR). We found that our data was not sufficient to study the Individual Hotspot Usage, which requires either a larger number of meioses per individual (Ma et al. 2015; Kadri et al. 2016a; Sandor et al. 2012a) or denser genotyping in families (Coop et al. 2008).

Our approach to study the genetic determinism of GRR in the Lacaune population was first to estimate its heritability, using a classical analysis in a large pedigree. This analysis also allowed to extract additive genetic values (EBVs) for the trait in 345 male parents, which we used for a GWAS in a second step. The EBVs are by definition, only determined by genetic factors, as environmental effects on GRR are averaged out. Indeed, we found that the proportion of variance in EBVs explained by genetic factors in the GWAS was essentially one. A consequence was that, although this sample size could be deemed low in current standards, the power of our GWAS was greatly increased by the high precision on the phenotype. We estimated the heritability of GRR at 0.23, which was similar to estimates from studies on the same phenotype in ruminants (*e.g.* 0.22 in cattle (Sandor et al. 2012a) or 0.12 in male Soay sheep (Susan E. Johnston et al. 2016), but see below for

a discussion on the comparison with Soay sheep). We had little information on the environmental factors that could influence recombination rate, but did find a suggestive effect of the month of insemination on GRR, especially we found increased GRR at the month of May. Confirmation and biological interpretation of this result would need dedicated studies, but it was consistent with the fact that fresh (*i.e.* not frozen) semen is used for insemination in sheep and that the reproduction of this species is seasonal (Rosa and Bryant 2003).

The genetic determinism of GRR discovered in our study closely resembles what has been found in previous studies, especially in mammals. Two major loci and two suggestive ones affected recombination rate in Lacaune. The two main QTLs are common to cattle and Soay sheep. The underlying genes and mutations for these two QTLs are not yet resolved but the fact that the two regions harbour interacting genes (*RNF212* and *HEI10* (Qiao et al. 2014; Rao et al. 2016)) involved in the maturation of crossovers, make these two genes likely functional candidates. Indeed, these two genes were identified as potential candidates underlying QTLs for GRR in mice (R. J. Wang and Payseur 2017). The third gene identified here, *KCNJ15*, is a novel candidate, and its role and mechanism of action in the repair of DSBs needs to be confirmed and elucidated. Interestingly, these three genes are linked to the reparation of DSBs and crossover maturation processes. Finally, the fourth candidate *FSHR* has well documented effects on gametogenesis but has not been linked to recombination previously.

In our study, sixty percent of the additive genetic variance in GRR remained unexplained by large effect QTLs and were due to polygenic effects. This could be interpreted in the light of recent evidence that has shown that other mechanisms, involved in chromosome conformation during meiosis, explain a substantial part of the variation in recombination rate between mouse strains (Baier et al. 2014) and bovids (Ruiz-Herrera et al. 2017). Furthermore the variations at the major mammal recombination loci (*RNF212*, *CPLX1*, *REC8* or the Human inversion *17q21.31*) explain only 3 to 11% (Ritz, Noor, and Singh 2017) of the phenotypic variance among individuals. Elucidating the genetic determinism of these different processes would thus require much larger sample sizes or different experimental approaches (Baier et al. 2014; Ruiz-Herrera et al. 2017).

The combination of datasets from the Lacaune population and one from the recent study of recombination in Soay sheep (Susan E. Johnston et al. 2016) allowed to study the evolution of recombination at relatively short time scales. One of the most striking difference between our two

studies is that the two QTLs that were detected in common had no effect in Soay males, whereas they had strong effects in Lacaune males. However, the two populations had very similar polygenic heritability: accounting for the fact that the Lacaune QTLs explain about 40% of the additive genetic variance, we could estimate the polygenic additive genetic variance in Lacaune males at 0.16, very similar to the 0.12 found in Soay males. Combined with our results that the two populations exhibit very similar male recombination maps, both in terms of intensity and genome distribution, the combination of the two studies shows that recombination patterns are conserved between populations under distinct genetic determinism, highlighting the robustness of mechanisms that drive them. Further work is needed to get a more detailed picture of the genetic control of recombination in sheep and will likely require combining multiple inferences from genetics, cytogenetics, molecular biology and bioinformatics analyses.

Acknowledgments

Institut de l'Elevage (JM. Astruc) and breed organizations (Ovitest and Confédération de Roquefort) provided the SNP genotypes and pedigree information. This work was partially funded by the BoDeliRe grant of the INRA Selgen Metaprogram and by Région Midi-Pyrénées. We are thankful to Tom Druet, Laurent Duret, Alain Pinton and Pierre Sourdille for their helpful comments on the manuscript.

Literature Cited

Ahlawat, S., P. Sharma, R. Sharma, R. Arora, N. K. Verma, B. Brahma, P. Mishra, and S. De. 2016. "Evidence of Positive Selection and Concerted Evolution in the Rapidly Evolving PRDM9 Zinc Finger Domain in Goats and Sheep." *Animal Genetics* 47 (6): 740–51.

Auton, Adam, Ying Rui Li, Jeffrey Kidd, Kyle Oliveira, Julie Nadel, J. Kim Holloway, Jessica J. Hayward, et al. 2013. "Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs." *PLoS Genetics* 9 (12): e1003984.

Baier, Brian, Patricia Hunt, Karl W. Broman, and Terry Hassold. 2014. "Variation in Genome-Wide Levels of Meiotic Recombination Is Established at the Onset of Prophase in Mammalian Males." *PLoS Genetics* 10 (1): e1004125.

Baloche, G., A. Legarra, G. Sallé, H. Larroque, J-M Astruc, C. Robert-Granié, and F. Barillet. 2014. "Assessment of Accuracy of Genomic Prediction for French Lacaune Dairy Sheep." *Journal of Dairy Science* 97 (2): 1107–16.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48.

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. 2010. "PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice." *Science* 327 (5967): 836–40.

Berg, Ingrid L., Rita Neumann, Kwan-Wood G. Lam, Shriparna Sarbajna, Linda Odenthal-Hesse, Celia A. May, and Alec J. Jeffreys. 2010. "PRDM9 Variation Strongly Influences Recombination Hot-Spot Activity and Meiotic Instability in Humans." *Nature Genetics* 42 (10): 859–63.

Berg, Ingrid L., Rita Neumann, Shriparna Sarbajna, Linda Odenthal-Hesse, Nicola J. Butler, and Alec J. Jeffreys. 2011. "Variants of the Protein PRDM9 Differentially Regulate a Set of Human Meiotic Recombination Hotspots Highly Active in African Populations." *Proceedings of the National Academy of Sciences of the United States of America* 108 (30): 12378–83.

Boitard, Simon, Willy Rodríguez, Flora Jay, Stefano Mona, and Frédéric Austerlitz. 2016. "Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach." *PLoS Genetics* 12 (3): e1005877.

Brunschwig, Hadassa, Liat Levi, Eyal Ben-David, Robert W. Williams, Benjamin Yakir, and Sagiv Shifman. 2012. "Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome." *Genetics* 191 (3): 757–64.

Chan, Andrew H., Paul A. Jenkins, and Yun S. Song. 2012. "Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila Melanogaster." *PLoS Genetics* 8 (12): e1003090.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (February): 7.

Cheung, Vivian G., Joshua T. Burdick, Deborah Hirschmann, and Michael Morley. 2007. "Polymorphic Variation in Human Meiotic Recombination." *American Journal of Human Genetics* 80 (3): 526–30.

Chowdhury, Reshmi, Philippe R. J. Bois, Eleanor Feingold, Stephanie L. Sherman, and Vivian G. Cheung. 2009. "Genetic Analysis of Variation in Human Meiotic Recombination." *PLoS Genetics* 5 (9): e1000648.

Cirulli, Elizabeth T., Richard M. Kliman, and Mohamed A. F. Noor. 2007. "Fine-Scale Crossover Rate Heterogeneity in Drosophila Pseudoobscura." *Journal of Molecular Evolution* 64 (1): 129–35.

Cohen-Zinder, Miri, Eyal Seroussi, Denis M. Larkin, Juan J. Loor, Annelie Everts-van der Wind, Jun-Heon Lee, James K. Drackley, et al. 2005. "Identification of a Missense Mutation in the Bovine ABCG2 Gene with a Major Effect on the QTL on Chromosome 6 Affecting Milk Yield and Composition in Holstein Cattle." *Genome Research* 15 (7): 936–44.

Coop, Graham, Xiaoquan Wen, Carole Ober, Jonathan K. Pritchard, and Molly Przeworski. 2008. "High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns among Humans." *Science* 319 (5868): 1395–98.

Cox, Allison, Cheryl L. Ackert-Bicknell, Beth L. Dumont, Yueming Ding, Jordana Tzenova Bell, Gudrun A. Brockmann, Jon E. Wergedal, et al. 2009. “A New Standard Genetic Map for the Laboratory Mouse.” *Genetics* 182 (4): 1335–44.

Crawford, Dana C., Tushar Bhangale, Na Li, Garrett Hellenthal, Mark J. Rieder, Deborah A. Nickerson, and Matthew Stephens. 2004. “Evidence for Substantial Fine-Scale Variation in Recombination Rates across the Human Genome.” *Nature Genetics* 36 (7). Nature Publishing Group: 700–706.

Druet, Tom, and Michel Georges. 2015. “LINKPHASE3: An Improved Pedigree-Based Phasing Algorithm Robust to Genotyping and Map Errors.” *Bioinformatics* 31 (10): 1677–79.

Fariello, Maria-Ines, Bertrand Servin, Gwenola Tosser-Klopp, Rachel Rupp, Carole Moreno, International Sheep Genomics Consortium, Magali San Cristobal, and Simon Boitard. 2014. “Selection Signatures in Worldwide Sheep Populations.” *PloS One* 9 (8): e103813.

Ferrer-Admetlla, Anna, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. 2014. “On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure.” *Molecular Biology and Evolution* 31 (5): 1275–91.

Fraley, Chris, and Adrian E. Raftery. 2002. “Model-Based Clustering, Discriminant Analysis, and Density Estimation.” *Journal of the American Statistical Association* 97 (458). Taylor & Francis: 611–31.

Groenen, Martien A. M., Alan L. Archibald, Hirohide Uenishi, Christopher K. Tuggle, Yasuhiro Takeuchi, Max F. Rothschild, Claire Rogel-Gaillard, et al. 2012. “Analyses of Pig Genomes Provide Insight into Porcine Demography and Evolution.” *Nature* 491 (7424): 393–98.

Groenen, Martien A. M., Per Wahlberg, Mario Foglio, Hans H. Cheng, Hendrik-Jan Megens, Richard P. M. A. Crooijmans, Francois Besnier, et al. 2009. “A High-Density SNP-Based Linkage Map of the Chicken Genome Reveals Sequence Features Correlated with Recombination Rate.” *Genome Research* 19 (3): 510–19.

Guan, Yongtao, and Matthew Stephens. 2008. “Practical Issues in Imputation-Based Association Mapping.” *PLoS Genetics* 4 (12): e1000279.

Hassold, Terry, Heather Hall, and Patricia Hunt. 2007. “The Origin of Human Aneuploidy: Where We Have Been, Where We Are Going.” *Human Molecular Genetics* 16 (R2): R203–8.

Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. “A Flexible and Accurate Genotype Imputation Method for the next Generation of Genome-Wide Association Studies.” *PLoS Genetics* 5 (6): e1000529.

Ignatieva, Elena V., Victor G. Levitsky, Nikolay S. Yudin, Mikhail P. Moshkin, and Nikolay A. Kolchanov. 2014. “Genetic Basis of Olfactory Cognition: Extremely High Level of DNA Sequence Polymorphism in Promoter Regions of the Human Olfactory Receptor Genes Revealed Using the 1000 Genomes Project Dataset.” *Frontiers in Psychology* 5 (March): 247.

International HapMap Consortium, Kelly A. Frazer, Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, et al. 2007. “A Second Generation Human Haplotype Map of over 3.1 Million SNPs.” *Nature* 449 (7164): 851–61.

Jiang, Yu, Min Xie, Wenbin Chen, Richard Talbot, Jillian F. Maddox, Thomas Faraut, Chunhua Wu, et al. 2014. “The Sheep Genome Illuminates Biology of the Rumen and Lipid Metabolism.” *Science* 344 (6188): 1168–73.

Johnston, S. E., J. Huisman, P. A. Ellis, and J. M. Pemberton. 2017. “A High-Density Linkage Map Reveals Sexually-Dimorphic Recombination Landscapes in Red Deer (Cervus Elaphus).” *bioRxiv*. biorxiv.org. http://biorxiv.org/content/early/2017/01/13/100131.abstract.

Johnston, Susan E., Camillo Bérénos, Jon Slate, and Josephine M. Pemberton. 2016. “Conserved Genetic Architecture Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep (Ovis Aries).” *Genetics* 203 (1): 583–98.

Johnston, Susan E., Jacob Gratten, Camillo Berenos, Jill G. Pilkington, Tim H. Clutton-Brock, Josephine M. Pemberton, and Jon Slate. 2013. “Life History Trade-Offs at a Single Locus Maintain Sexually Selected Genetic Variation.” *Nature* 502 (7469): 93–95.

Kadri, Naveen Kumar, Chad Harland, Pierre Faux, Nadine Cambisano, Latifa Karim, Wouter Coppieters, Sébastien Fritz, et al. 2016a. “Coding and Noncoding Variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B Affect Recombination Rate in Cattle.” *Genome Research* 26 (10): 1323–32.

———. 2016b. “Coding and Noncoding Variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B

Affect Recombination Rate in Cattle." *Genome Research* 26 (10): 1323–32.
Kari, Vijayalakshmi, Andrei Shchebet, Heinz Neumann, and Steven A. Johnsen. 2011. "The H2B Ubiquitin Ligase RNF40 Cooperates with SUPT16H to Induce Dynamic Changes in Chromatin Structure during DNA Double-Strand Break Repair." *Cell Cycle* 10 (20): 3495–3504.
Kaur, Taniya, and Matthew V. Rockman. 2014. "Crossover Heterogeneity in the Absence of Hotspots in Caenorhabditis Elegans." *Genetics* 196 (1): 137–48.
Kijas, James, J. Lenstra, B. Hayes, S. Boitard, Porto Neto L, Magali San Cristobal, Bertrand Servin, et al. 2012. "Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection." *PLoS Biology*. journals.plos.org. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001258.
Kim, E-S, A. R. Elbeltagy, A. M. Aboul-Naga, B. Rischkowsky, B. Sayre, J. M. Mwacharo, and M. F. Rothschild. 2016. "Multiple Genomic Signatures of Selection in Goats and Sheep Indigenous to a Hot Arid Environment." *Heredity* 116 (3): 255–64.
Kong, A., G. Thorleifsson, H. Stefansson, G. Masson, A. Helgason, D. F. Gudbjartsson, G. M. Jonsdottir, et al. 2008a. "Sequence Variants in the RNF212 Gene Associate with Genome-Wide Recombination Rate." *Science* 319 (5868): 1398–1401.
———. 2008b. "Sequence Variants in the RNF212 Gene Associate with Genome-Wide Recombination Rate." *Science* 319 (5868): 1398–1401.
Kong, Augustine, Gudmar Thorleifsson, Michael L. Frigge, Gisli Masson, Daniel F. Gudbjartsson, Rasmus Villemoes, Erna Magnusdottir, Stefania B. Olafsdottir, Unnur Thorsteinsdottir, and Kari Stefansson. 2014. "Common and Low-Frequency Variants Associated with Genome-Wide Recombination Rate." *Nature Genetics* 46 (1): 11–16.
Kong, Augustine, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G. Bragi Walters, et al. 2010. "Fine-Scale Recombination Rate Differences between Sexes, Populations and Individuals." *Nature* 467 (7319): 1099–1103.
Lange, Julian, Shintaro Yamada, Sam E. Tischfield, Jing Pan, Seoyoung Kim, Xuan Zhu, Nicholas D. Socci, Maria Jasin, and Scott Keeney. 2016. "The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair." *Cell* 167 (3): 695–708.e16.
Li, Heng, and Richard Durbin. 2011. "Inference of Human Population History from Individual Whole-Genome Sequences." *Nature* 475 (7357): 493–96.
Li, Na, and Matthew Stephens. 2003. "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data." *Genetics* 165 (4): 2213–33.
Ma, Li, Jeffrey R. O'Connell, Paul M. VanRaden, Botong Shen, Abinash Padhi, Chuanyu Sun, Derek M. Bickhart, et al. 2015. "Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis." *PLoS Genetics* 11 (11): e1005387.
Mancera, Eugenio, Richard Bourgon, Alessandro Brozzi, Wolfgang Huber, and Lars M. Steinmetz. 2008. "High-Resolution Mapping of Meiotic Crossovers and Non-Crossovers in Yeast." *Nature* 454 (7203): 479–85.
McVean, Gil, Philip Awadalla, and Paul Fearnhead. 2002. "A Coalescent-Based Method for Detecting and Estimating Recombination from Gene Sequences." *Genetics* 160 (3): 1231–41.
Mihola, Ondrej, Zdenek Trachtulec, Cestmir Vlcek, John C. Schimenti, and Jiri Forejt. 2009. "A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase." *Science* 323 (5912): 373–75.
Moreno-Romieux, Carole, Flavie Tortereau, Jérome Raoul, and Bertrand Servin. 2017. "High Density Genotypes of French Sheep Populations." doi:10.5281/zenodo.237116.
Myers, Simon, Leonardo Bottolo, Colin Freeman, Gil McVean, and Peter Donnelly. 2005. "A Fine-Scale Map of Recombination Rates and Hotspots across the Human Genome." *Science* 310 (5746): 321–24.
Myers, S., C. C. A. Spencer, A. Auton, L. Bottolo, C. Freeman, P. Donnelly, and G. McVean. 2006. "The Distribution and Causes of Meiotic Recombination in the Human Genome." *Biochemical Society Transactions* 34 (Pt 4): 526–30.
Nagamine, Yoshitaka, Ricardo Pong-Wong, Pau Navarro, Veronique Vitart, Caroline Hayward, Igor Rudan, Harry Campbell, et al. 2012. "Localising Loci Underlying Complex Trait Variation Using Regional Genomic Relationship Mapping." *PloS One* 7 (10). Public Library of Science: e46501.
Nagel, Anja C., Patrick Fischer, Jutta Szawinski, Martina K. La Rosa, and Anette Preiss. 2012. "Cyclin G Is

Involved in Meiotic Recombination Repair in Drosophila Melanogaster." *Journal of Cell Science* 125 (Pt 22): 5555–63.
Norris, Belinda J., and Vicki A. Whan. 2008. "A Gene Duplication Affecting Expression of the Ovine ASIP Gene Is Responsible for White and Black Sheep." *Genome Research* 18 (8): 1282–93.
O'Reilly, Paul F., Ewan Birney, and David J. Balding. 2008. "Confounding between Recombination and Selection, and the Ped/Pop Method for Detecting Selection." *Genome Research* 18 (8): 1304–13.
Pratto, Florencia, Kevin Brick, Pavel Khil, Fatima Smagulova, Galina V. Petukhova, and R. Daniel Camerini-Otero. 2014. "Recombination Initiation Maps of Individual Human Genomes." *Science* 346 (6211). American Association for the Advancement of Science: 1256442.
Qiao, Huanyu, H. B. D. Prasada Rao, Ye Yang, Jared H. Fong, Jeffrey M. Cloutier, Dekker C. Deacon, Kathryn E. Nagel, et al. 2014. "Antagonistic Roles of Ubiquitin Ligase HEI10 and SUMO Ligase RNF212 Regulate Meiotic Recombination." *Nature Genetics* 46 (2): 194–99.
Rao, H. B. D. Prasada, Huanyu Qiao, Shubhang K. Bhatt, Logan R. J. Bailey, Hung D. Tran, Sarah L. Bourne, Wendy Qiu, et al. 2016. "A SUMO-Ubiquitin Relay Recruits Proteasomes to Chromosome Axes to Regulate Meiotic Recombination." *bioRxiv*. Cold Spring Harbor Labs Journals. doi:10.1101/095711.
Reynolds, April, Huanyu Qiao, Ye Yang, Jefferson K. Chen, Neil Jackson, Kajal Biswas, J. Kim Holloway, et al. 2013. "RNF212 Is a Dosage-Sensitive Regulator of Crossing-over during Mammalian Meiosis." *Nature Genetics* 45 (3): 269–78.
Ritz, Kathryn R., Mohamed A. F. Noor, and Nadia D. Singh. 2017. "Variation in Recombination Rate: Adaptive or Not?" *Trends in Genetics: TIG* 33 (5): 364–74.
Rochus, Christina Marie, Flavie Tortereau, Florence Plisson-Petit, Gwendal Restoux, Carole Moreno-Romieux, Gwenola Tosser-Klopp, and Bertrand Servin. 2017. "High Density Genome Scan for Selection Signatures in French Sheep Reveals Allelic Heterogeneity and Introgression at Adaptive Loci." *bioRxiv*. Cold Spring Harbor Labs Journals. doi:10.1101/103010.
Rockman, Matthew V., and Leonid Kruglyak. 2009. "Recombinational Landscape and Population Genomics of Caenorhabditis Elegans." *PLoS Genetics* 5 (3): e1000419.
Rosa, H. J. D., and M. J. Bryant. 2003. "Seasonality of Reproduction in Sheep." *Small Ruminant Research: The Journal of the International Goat Association* 48 (3): 155–71.
Ruiz-Herrera, Aurora, Miluse Vozdova, Jonathan Fernández, Hana Sebestova, Laia Capilla, Jan Frohlich, Covadonga Vara, et al. 2017. "Recombination Correlates with Synaptonemal Complex Length and Chromatin Loop Size in Bovids-Insights into Mammalian Meiotic Chromosomal Organization." *Chromosoma*, January. doi:10.1007/s00412-016-0624-3.
Sabeti, Pardis C., David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, et al. 2002. "Detecting Recent Positive Selection in the Human Genome from Haplotype Structure." *Nature* 419 (6909): 832–37.
Sandor, Cynthia, Wanbo Li, Wouter Coppieters, Tom Druet, Carole Charlier, and Michel Georges. 2012a. "Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle." *PLoS Genetics* 8 (7): e1002854.
———. 2012b. "Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle." *PLoS Genetics* 8 (7): e1002854.
Scheet, Paul, and Matthew Stephens. 2006. "A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase." *American Journal of Human Genetics* 78 (4): 629–44.
Servin, Bertrand, and Matthew Stephens. 2007. "Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits." *PLoS Genetics* 3 (7): e114.
Shifman, Sagiv, Jordana Tzenova Bell, Richard R. Copley, Martin S. Taylor, Robert W. Williams, Richard Mott, and Jonathan Flint. 2006. "A High-Resolution Single Nucleotide Polymorphism Genetic Map of the Mouse Genome." *PLoS Biology* 4 (12): e395.
Słabicki, Mikołaj, Mirko Theis, Dragomir B. Krastev, Sergey Samsonov, Emeline Mundwiller, Magno Junqueira, Maciej Paszkowski-Rogacz, et al. 2010. "A Genome-Scale DNA Repair RNAi Screen Identifies SPG48 as a Novel Gene Associated with Hereditary Spastic Paraplegia." *PLoS Biology* 8 (6): e1000408.

Stathopoulos, Sofia, Jacqueline M. Bishop, and Colleen O'Ryan. 2014. "Genetic Signatures for Enhanced Olfaction in the African Mole-Rats." *PloS One* 9 (4): e93336.
Stephens, Matthew. 2017. "False Discovery Rates: A New Deal." *Biostatistics* 18 (2): 275–94.
Stevison, Laurie S., and Mohamed A. F. Noor. 2010. "Genetic and Evolutionary Correlates of Fine-Scale Recombination Rate Variation in Drosophila Persimilis." *Journal of Molecular Evolution* 71 (5-6): 332–45.
Storey, J. D., and R. Tibshirani. 2003. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences* 100 (16): 9440–45.
Sturtevant, A. H. 1913. "The Linear Arrangement of Six Sex-Linked Factors in Drosophila, as Shown by Their Mode of Association." *The Journal of Experimental Zoology* 14 (1). Wiley Subscription Services, Inc., A Wiley Company: 43–59.
Takasuga, Akiko. 2015. "PLAG1 and NCAPG-LCORL in Livestock." *Animal Science Journal = Nihon Chikusan Gakkaiho*. Wiley Online Library. http://onlinelibrary.wiley.com/doi/10.1111/asj.12417/pdf.
Tapanainen, J. S., K. Aittomäki, J. Min, T. Vaskivuo, and I. T. Huhtaniemi. 1997. "Men Homozygous for an Inactivating Mutation of the Follicle-Stimulating Hormone (FSH) Receptor Gene Present Variable Suppression of Spermatogenesis and Fertility." *Nature Genetics* 15 (2): 205–6.
Tortereau, Flavie, Bertrand Servin, Laurent Frantz, Hendrik-Jan Megens, Denis Milan, Gary Rohrer, Ralph Wiedmann, et al. 2012. "A High Density Recombination Map of the Pig Reveals a Correlation between Sex-Specific Recombination and GC Content." *BMC Genomics* 13 (November): 586.
Voight, Benjamin F., Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K. Pritchard. 2006. "A Map of Recent Positive Selection in the Human Genome." *PLoS Biology* 4 (3): e72.
Wang, Jianbin, H. Christina Fan, Barry Behr, and Stephen R. Quake. 2012. "Genome-Wide Single-Cell Analysis of Recombination Activity and de Novo Mutation Rates in Human Sperm." *Cell* 150 (2): 402–12.
Wang, Richard J., and Bret A. Payseur. 2017. "Genetics of Genome-Wide Recombination Rate Evolution in Mice from an Isolated Island." *Genetics*, June. doi:10.1534/genetics.117.202382.
Zeileis, Achim, and Gabor Grothendieck. 2005. "Zoo: S3 Infrastructure for Regular and Irregular Time Series." *Journal of Statistical Software, Articles* 14 (6): 1–27.
Zhou, Xiang, Peter Carbonetto, and Matthew Stephens. 2013. "Polygenic Modeling with Bayesian Sparse Linear Mixed Models." *PLoS Genetics* 9 (2): e1003264.

Annexe 2

Formulaire de demande de crédits incitatifs 2016 - Département de Génétique Animale

Titre du projet: Etude fine de la recombinaison chez la race admixée Romane
Champ thématique: CT1
Acronyme : Romane Ite Domum
Nom du porteur : Bertrand Servin (Genphyse, Dynagen)
Partenaire 2: Dominique Hazard, Carole Moreno, Flavie Tortereau (Genphyse, GeSPR)
Partenaire 3: Stéphane Fabre (Genphyse, GenRoC)
Pateforme impliquée: LaboGena

Contexte et état de l'art

La recombinaison est un processus biologique fondamental dont les caractéristiques fines et le déterminisme génétique peuvent être étudiés grâce aux données de génotypage dense. En ce qui concerne les mammifères, les premières études génomiques de la recombinaison ont été effectuées chez l'homme et la souris il y une dizaine d'années et ont démontré plusieurs phénomènes importants:

1. Les crossing-overs ne se répartissent pas de façon homogène sur le génome, en particulier à l'échelle de la kilobase. De petites régions génomiques nommées points chauds de recombinaison sont extrêmement enrichies en crossing overs lors des méioses.
2. Il est possible de caractériser la variabilité individuelle dans le processus de recombinaison, d'une part de par son intensité (nombre de crossing over par méiose soit le Taux de Recombinaison Total, TRT) et d'autre part de par sa répartition sur le génome (Biais d'Usage des Points Chauds, BUPC). A partir de cette caractérisation il est possible de démontrer que la variabilité de ces deux phénotypes est soumise à un contrôle génétique et de détecter des QTLs de recombinaison pour le TRT et le BUPC.

En terme de méthodologie, les études de la recombinaison sur données génomiques denses peuvent exploiter différents types de données:

1. Des **données familiales,** où les descendants et les parents sont génotypés pour des puces SNPs denses. Chez le bovin des données de puce 60K ont été utilisées efficacement dans ce cadre (*e.g.* Sandor et al. 2012)
2. Des **données populationnelles** d'individus non apparentés génotypés pour des marqueurs très denses. L'idée est d'exploiter les patrons de Déséquilibre de Liaison (DL) observés pour en déduire un taux de recombinaison historique modélisé comme ayant conduit au DL actuel. C'était jusqu'à récemment la seule approche permettant d'identifier précisément les points chauds de recombinaison à l'échelle d'un génome entier (*e.g.* Myers et al. 2007).
3. Des données chez des **individus admixés** entre plusieurs populations, par exemple les populations afro-américaines chez l'homme (Wegmann et al. 2011). L'idée ici est que les chromosomes de ces individus sont des mosaïques de chromosomes ancestraux issus des populations d'origine. Il est possible de reconstituer ces mosaïques à partir de données de génotypage dense. Une fois cette reconstruction effectuée, on peut exploiter la distribution des jonctions entre segments d'admixture pour estimer le taux de recombinaison à l'échelle du génome.

Chez les animaux d'élevage les premières études génomiques de la recombinaison ont commencé, en particulier chez le bovin (*e.g.* Sandor et al. 2012). Ces données exploitent les informations obtenues dans le cadre de la sélection génomique où de grandes familles sont génotypées pour une puce SNP de 60,000 marqueurs environ.

Au sein du laboratoire, dans les équipes GeSPR, GenRoc et DynaGen nous avons entrepris dans le projet Selgen BoDeliRe d'étudier le processus de recombinaison chez le mouton, dans le cadre de la thèse de Morgane Petit. Ce travail a permis de décrire plusieurs phénomènes importants. Nous avons aujourd'hui pu exploiter les informations disponibles en race Lacaune et démontré le déterminisme génétique du TRT, détecté deux QTLs d'effet important ségrégeant dans la population. En parallèle, nous avons pu exploiter des données populationnelles de génotypage haute-densité (600K SNP) pour décrire de nombreux points-chauds de recombinaison dans la race Lacaune. En revanche nous n'avons pas réussi à étudier de manière satisfaisante la variabilité individuelle du BUPC dans la race Lacaune.

Nos conclusions sont que nous ne disposons pas de données suffisantes pour le faire. Une solution consiste à augmenter la densité de marquage dans des familles pour pouvoir améliorer notre précision de localisation des crossing-overs (Coop et al. 2010). Il existe une opportunité d'exploiter cette approche en race Romane qui permettrait de complémenter de manière très intéressante les travaux déjà entrepris.

Objectifs

Dans le cadre du projet *Romane Ite Domum*, nous proposons d'augmenter le nombre de génotypages 600K dans de grandes familles de pères et nucléaires dans la race Romane. Ces familles sont d'ores et déja disponibles et le financement demandé couvrira une partie des frais de génotypage 600K. Par ailleurs, dans le cadre d'un autre projet (Julie Demars) des données préliminaires (78 animaux dans 5 familles de pères, soit 73 méioses) nous ont permis d'évaluer les résultats attendus. Ces nouvelles données nous permettront:

1. **de disposer d'une collection de crossing-overs bien résolus** (de l'ordre de la centaine de Kb cf. Figure 1) chez un petit nombre d'animaux. Ceci nous permettra d'établir un phénotypage fin des individus en terme de BUPC que nous chercherons à associer à un gène candidat majeur (PRDM9), nous proposons de nous concentrer sur ce gène spécifiquement.
2. **d'établir une carte de recombinaison basée sur les patrons d'admixture** à travers la reconstruction des nombreux haplotypes parentaux de race Romane (Figure 2).
3. A partir de la reconstruction des mosaïques dans les chromosomes Romane (Figure 2), **d'enrichir la collection d'haplotypes 600K disponibles dans les races d'origine Berrichon du Cher et Romanov** et ainsi de détecter les points chauds de recombinaison par une approche populationnelle dans ces deux races.

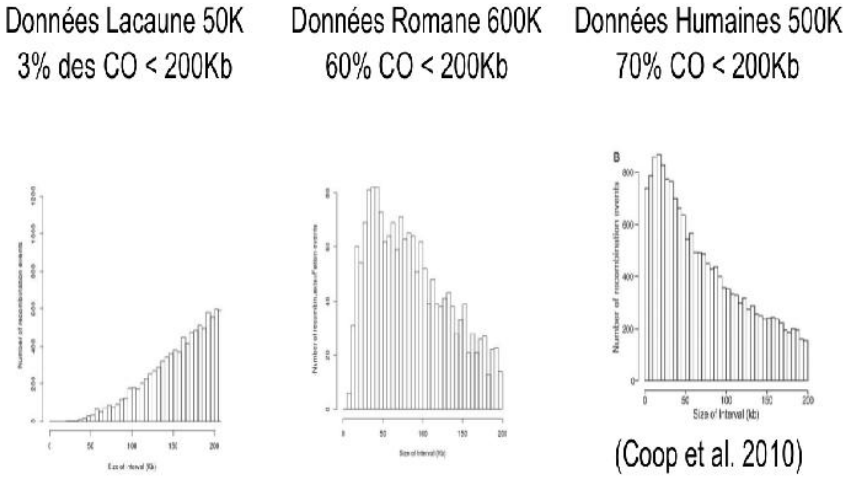

Figure 1 Distribution de la taille des intervalles de résolution des crossing-overs détectés dans 3 dispositifs.

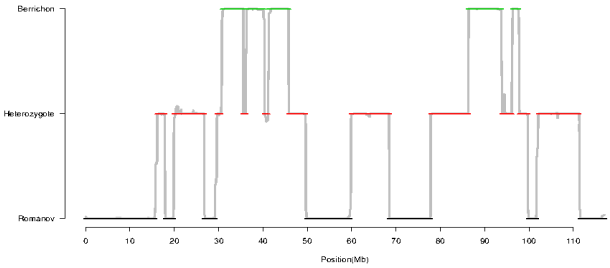

Figure 2 Exemple de reconstruction des origines populationnelles (Berrichon ou Romanov) des haplotypes d'un individu Romane génotypé avec la puce HD sur le chromosome 6 ovin.

Caractère novateur ou exploratoire du projet

Ce projet vient en complément d'une étude déjà réalisée en race Lacaune où des cartes populationnelles et familiales ont pu être établies mais dans laquelle les données ne sont pas suffisantes pour étudier la variabilité individuelle de la localisation des crossing overs (BPUC). Malgré ce lien évident avec nos travaux actuels, l'approche proposée ici est novatrice à plusieurs titres. Premièrement, une démarche combinant à la fois approche familiale, populationnelle et d'admixture sur un même jeu de données n'a jamais été mise en oeuvre à notre connaissance. Deuxièmement, l'étude de la variabilité de la répartition sur le génome des points chauds entre populations n'a pour l'instant été étudiée que chez l'Homme et la souris. Les races Berrichon, Lacaune et Romanov sont très éloignées génétiquement, correspondant à des origines génétiques ayant divergé il y a longtemps. Il est donc possible que nous mettions en évidence des différences de position entre points chauds dans ces races.

Plan de travail et calendrier

Dispositif Expérimental

Pour mettre en place le dispositif expérimental, nous nous sommes appuyés d'une part sur les analyses préliminaires des données disponibles et d'autre part sur le recensement des familles existantes.

Les analyses préliminaires montrent que la résolution attendue des crossing overs est du même ordre de grandeur que les études effectuées chez l'homme (figure 1). Cependant, nous pouvons constater un déficit dans le nombre de crossing overs résolus dans les intervalles de moins de 20Kb avec les familles de demi-frères disponibles. Nous proposons d'améliorer cette résolution en génotypant dans chaque famille en plus du père et de 5 descendants demi-frères, un sixième descendant, plein-frère d'un des 5 premiers ainsi que leur mère. Nous nous attendons ainsi à améliorer fortement notre puissance de résolution des phases paternelles et conséquemment les résolutions des crossing-overs. Par ailleurs ceci nous permettra d'étudier l'effet du sexe sur le TRT, effet qui a été démontré dans d'autres espèces.
Nous avons identifié deux protocoles expérimentaux déjà existants pour lesquels nous disposons d'échantillons d'ADN ou de sang dans des familles adaptées. Il s'agit d'une part de lignées divergentes Romane sur le caractère d'efficacité alimentaire, pour lesquelles nous proposons de génotyper les 10 pères fondateurs ainsi que leurs descendants directs. Ces 10 béliers ne sont pas apparentés au niveau de leur parents ou grands-parents. D'autre part nous disposons de deux protocoles de lignées divergentes pour la réactivité à l'homme et la réactivité sociale entre pairs. Au

sein de ces 4 lignées, nous pouvons une vingtaine de familles dont les pères ne sont pas apparentés au niveau de leur parents, grands-parents ou arrière grands-parents. Pour l'ensemble de ces protocoles, l'acquisition de données génétiques haute-densité favorisera grandement la détection de QTL et les études d'association avec les caractères sélectionnés.

Calendrier

Le projet se déroulera sur 8 mois, dont 4 à 5 mois prévus pour l'analyse statistique proprement dite et 2 à 3 mois prévus pour l'obtention des génotypes et la mise en place en amont des programmes d'analyse.

1. Génotypage de 30 familles Romane sur la puce Illumina Ovine HD 600K à Labogena. Pendant le délai nécessaire au génotypage mise en place des scripts d'analyse sur les données d'ores et déjà disponibles.
2. Nettoyage des données (filtres sur les taux de données manquantes, erreurs mendéliennes …) (1 Semaine)
3. Phasage et identification des crossing-overs avec le logiciel Linkphase (1 Semaine)
4. Estimation des cartes de recombinaison familiales, étude de l'effet sexe sur le TRT. (2 Semaines)
5. Reconstruction des origines populationnelles des haplotypes parentaux (logiciel elai) et établissement des cartes de recombinaisons basées sur les patrons d'admixture. (3 Semaines)
6. Etablissement des cartes de recombinaison populationnelles en Berrichon et Romanov. Etude de l'éventuelle différence de répartition des points chauds entre les races Berrichon, Romanov et Lacaune (les cartes Lacaune sont déjà disponibles). (1 mois)
7. Calcul des phénotypes de BUPC sur les parents du dispositif. Etude de l'association de ce phénotype avec les haplotypes au gène PRDM9. (1 mois)

Pour la tâche 1, la collecte des familles et des échantillons de sang ou d'ADN seront effectués par Flavie Tortereau et Dominique Hazard de l'équipe GesPR. La gestion des génotypages puces et du séquençage / génotypage du gène PRDM9 seront effectués par Stéphane Fabre et l'équipe GenROC. Pour les autres tâches, le travail d'analyse de données sera effectué majoritairement par Morgane Petit, encadrée par Bertrand Servin et Carole Moreno, dans le cadre de sa thèse. Pour les tâches 2 et 3, des programmes d'analyse sont d'ores et déjà disponibles permettant un travail rapide. Pour les tâches 4 et 5 une partie des scripts d'analyse sont disponibles ou seront développés dans l'attente des génotypages. Pour la tâche 6 (calcul du BUPC) et une partie de la tâche 5 (comparaison des cartes populationnelles entre races), un travail méthodologique sera nécessaire pour implémenter les méthodes d'analyse existant dans la littérature ou en imaginer de nouvelles. Si l'analyse bibliographique sera effectuée en amont, le développement méthodologique nécessite d'avoir les résultats des tâches précédentes pour commencer.

Résultats déjà acquis et résultats complémentaires attendus

Avec le type de dispositif expérimental prévu, chaque famille permettra d'étudier 8 méioses (6 méioses mâles et 2 méioses femelles), soit environ 300 crossing-overs méiotiques. Par ailleurs, la reconstruction des haplotypes parentaux fournira 8 haplotypes Romane indépendants par famille (2 du père, 2 de la mère et 4 haplotypes maternels transmis aux demi-frères restants). Sur la base des

données préliminaires nous avons estimé qu'un haplotype Romane permet d'identifier de l'ordre de 10 fois plus de crossing-overs historiques qu'une méiose. Chaque famille permettra donc d'identifier de l'ordre de 3000 crossing-overs historiques. Au final, sur la base du génotypage de 30 familles nous nous attendons à obtenir:

1. **9000 crossing-overs méiotiques,** répartis en 180 (6x30) méioses mâles et 60 (2x30) méioses femelles. Ceci permettra d'estimer l'effet du sexe sur le taux de recombinaison.
2. **Pour chaque père,** nous disposerons d'environ **200 crossing-overs méiotiques** dont 70% (140) sont attendus comme suffisamment bien résolus (< 200Kb) pour établir le phénotype de BUPC. Pour les mères, seulement de l'ordre de 70 à 80 crossing-overs seront disponibles, soit une cinquantaine pour l'analyse du BUPC.
3. **240 haplotypes Romane indépendants** permettant d'étudier les cartes de recombinaison populationnelle en Romanov et Berrichon (120 de chaque population sont attendus à une position donnée du génome).
4. De l'ordre de **100.000 crossing overs historiques** identifiés dans les haplotypes Romane, permettant d'établir des cartes de recombinaison sur la base de l'admixture. Sur la base de nos travaux en race Lacaune, pour laquelle nous disposons de 300.000 crossing-overs identifiés, nous nous attendons à pouvoir bien estimer les taux de recombinaison dans des fenêtres de l'ordre de la mégabase.

Bibliographie

Coop G, Wen Z, Ober C, Pritchard JK, Przeworksi M. (2008) High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. Science 31 Jan 2008 319(5868):1395-8

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. (2005) A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science 14 Oct 2005 : 321-324

Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR *et al.* **Recombination rates in admixed individuals identified by ancestry-based inference.** Nat Genet. 2011 Jul 20;43(9):847-53.

Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M. (2012) **Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle.** PLoS Genet 8(7): e1002854.

Justification budgétaire

Génotypage puce Haute-densité: 30 familles comprenant une mère un père et 6 descendants = 8 individus par famille * 30 familles = 240 puces HD * 170 euros = 40800 euros. Utilisation financement BoDeLire = 20800 euros. Demande département : 20000 euros.
Séquençage / Génotypages additionnels du gène PRDM9: 2000 euros.

AUTHOR : Morgane PETIT

TITLE : Study of the recombination patterns, of their genetic determinisms and of their impact on genomic selection in the ovine French breed Lacaune.

SUPERVISORS : Carole MORENO et Bertrand SERVIN

Summary: Genetic recombination is a fundamental biological process, which occurs during the meiosis. It allows the good segregation of the chromosomes and contributes to maintain the genetic diversity. Recombination was already studied in a lot of different species, especially in mammals and in farm animals, such as the pig, the cattle or the sheep. In each case, a variation of the recombination rate between the individuals was observed. This variation was heritable and under genetic determinism. In some species, genetic recombination maps were also created, which allowed to localize the crossovers and to detect really tiny genomic regions where the recombination is huge: the recombination hotspots. In the Lacaune breed sheep, a lot of genotyping data are available thanks to two existing arrays: a first with a medium density of markers (about 54,000 markers) and a second with a high density of markers (about 600,000 markers). Two datasets were thus available: a familial dataset with about 6,000 animals genotyped for the 54,000 markers and a dataset of 70 unrelated Lacaune genotyped for the 600,000 markers. Genetic recombination maps were created for these two datasets. With the 70 unrelated Lacaune, about 50,000 hotspots were detected. The familial dataset allowed to observe the mammals common recombination patterns. Finally, when the two datasets were combined, selection signatures were revealed and a high-density recombination map were created. Furthermore, a variation of the recombination rate within the individuals was observed and was associated to 2 main *QTLs* on the chromosomes 6 and 7. Already known, or not, candidate genes were proposed and sometimes studied: especially *RNF212* and *HEI10*. Finally, a comparison with another sheep breed revealed that the genetic recombination maps were really similar, but the individual recombination rate was under a different genetic determinism. A concrete application of the genetic recombination map in genomic selection was also proposed thanks to the creation of low-density *SNPs* sets, which could be used to impute the animals and thus to improve the genotyping and the genomic selection for lesser costs.

Key words: genetic recombination, crossovers, genetic recombination maps, hotspots, recombination rate variation, genetic determinism, Lacaune sheep.

ADMINISTRATIVE DISCIPLINE: Pathology, Toxicology, Genetic & Nutrition.

RESEARCH UNIT: INRA, UMR 1388 Génétique, Physiologie et Systèmes d'Elevages. 24 Chemin de Borde-Rouge, CS 52627, 31326 Castanet-Tolosan Cedex, France.

AUTEUR : Morgane PETIT

TITRE : Etude des patrons de recombinaison, de leur déterminisme génétique et de leur impact en sélection génomique en race ovine Lacaune.

DIRECTEURS DE THESE : Carole MORENO et Bertrand SERVIN

LIEU ET DATE DE SOUTENANCE : Toulouse, le 17 Octobre 2017.

Résumé : La recombinaison génétique est un processus biologique fondamental, ayant lieu au cours de la méiose et assurant la bonne ségrégation des chromosomes, ainsi que le maintien de la variabilité génétique grâce au brassage intrachromosomique. La recombinaison a été étudiée dans de nombreuses espèces, en particulier chez les Mammifères et les animaux d'élevage, comme les bovins, les porcs ou les ovins. Dans tous les cas, une variation du taux de recombinaison a été observée entre les individus et il a été démontré qu'elle était héritable et sous déterminisme génétique. Dans certaines espèces, des cartes génétiques ont également été construites, ce qui a permis de localiser les crossing-overs et de détecter de très petites zones du génome où la recombinaison était importante : les points chauds. En race ovine Lacaune, de nombreuses données de génotypages sont disponibles, notamment grâce à l'existence de deux puces : une de moyenne densité avec 54 000 marqueurs et une de haute densité avec 600 000 marqueurs. Deux jeux de données étaient donc disponibles ; un jeu de données familial avec près de 6 000 individus apparentés et génotypés pour les 54 000 marqueurs et un jeu de données comportant 70 Lacaune non apparentés et génotypés pour les 600 000 marqueurs. Des cartes génétiques ont donc été créées pour ces deux jeux de données. Avec les animaux non apparentés, environ 50 000 points chauds ont été détectés. Le jeu de données familial a permis d'observer des motifs de distribution de la recombinaison communs aux autres Mammifères. Enfin, la combinaison des deux jeux de données a révélé la présence de signatures de sélection et a permis de créer une carte génétique de haute densité. De plus, une variation du taux de recombinaison a été observée entre les individus et a pu être liée à l'existence de 2 *QTLs* majeurs sur les chromosomes 6 et 7. Des gènes candidats plus ou moins bien connus ont pu être proposés, voire étudiés : *RNF212* et *HEI10*. De plus, une comparaison avec une autre population ovine a permis de

montrer que les cartes de recombinaison étaient quasiment identiques, mais que le taux de recombinaison individuel était soumis à un déterminisme génétique différent. Il a également été possible de proposer une application concrète pour l'utilisation des cartes génétiques en sélection génomique, grâce à la création de puces basse densité pouvant être utilisées pour l'imputation des reproducteurs et donc favoriser le génotypage et la sélection génomique à moindre coût.

Mots-clés : recombinaison génétique, crossing-overs, cartes génétiques, points chauds, variation de la recombinaison, déterminisme génétique, ovin, Lacaune.

DISCIPLINE ADMINISTRATIVE : Pathologie, Toxicologie, Génétique & Nutrition.

UNITE DE RECHERCHE : INRA, UMR 1388 Génétique, Physiologie et Systèmes d'Elevages. 24 Chemin de Borde-Rouge, CS 52627, 31326 Castanet-Tolosan Cedex, France.