Contributions à la modélisation statistique de données génétiques

Bertrand Servin

Contributions à la modélisation statistique de données génétiques

Mémoire présenté pour l'obtention du diplôme d'Habilitation à Diriger les Recherches

Bertrand SERVIN

Chargé de Recherche, INRA
Laboratoire Génétique, Physiologie, Systèmes d'Elevage.

Université de Toulouse

He not busy being born is busy dying

Bob Dylan

Table des matières

Introduction

La génétique est une discipline qui propose des approches spécifiques pour étudier la biologie des organismes, une part des différences inter-individuelles des caractères étant due à leur variabilité génétique. Détecter les facteurs génétiques spécifiques (mutations, gènes) influençant ces différences permet donc de comprendre, en partie, les mécanismes biologiques qui participent à l'élaboration des caractères observables (phénotypes). Par ailleurs, comme la génétique concerne une information qui est transmise à travers les générations successives, *i.e.* l'hérédité, elle offre par nature une interprétation évolutive des phénomènes biologiques. Les statistiques et la génétique sont deux disciplines étroitement liées. En effet, au delà de l'intérêt des statistiques pour les sciences expérimentales de manière générale, la génétique a une dimension fondamentalement aléatoire du fait du mécanisme Mendélien de transmission des gènes lors de la méiose. Plus encore, les recherches en biologie évolutive et en génétique ont conduit à d'importantes contributions au développement de la théorie statistique depuis les travaux de Galton et Fisher. On peut par exemple citer le développement des équations de prédiction des modèles mixtes (Henderson *et al.*, 1959), le développement de méthodes pour le contrôle du False Discovery Rate (FDR) (Storey et Tibshirani, 2003 ; Stephens, 2017) ou le principe de l'allocation de Dirichlet latente (Pritchard *et al.*, 2000). Ces travaux ont pour origine des questions de recherche en génétique mais ont conduit à des développements statistiques utilisés beaucoup plus largement. Aujourd'hui le développement de la génomique a multiplié les opportunités de recherche en génétique statistique, domaine dans lequel mes travaux se situent.

Ma carrière de chercheur a commencé en l'an 2000 lors de ma dernière année d'école d'ingénieur agronome au cours de laquelle j'ai effectué mon stage de DEA avec Frédéric Hospital sur l'optimisation du backcross assisté par marqueurs. J'ai ensuite poursuivi une thèse sur le sujet plus large de l'optimisation de programmes de sélection assistée par marqueurs dans le contexte de la sélection végétale. Ce début de carrière a commencé au même moment que l'aire de la génomique. En effet, c'est à cette époque que les séquences complètes des premiers génomes humains ont été publiées (The International Human Genome Sequencing Consortium, 2001 ; Venter *et al.*, 2001), que le principe de la sélection génomique a été décrit (Meuwissen *et al.*, 2001) et que paraissait un modèle qui s'avèrera fondamental pour la génomique des populations (Pritchard *et al.*, 2000). Mon travail de thèse m'a donné le goût du développement de méthodes pour l'étude de problèmes de génétique et j'ai eu la chance d'obtenir un poste de post-doctorant au sein du département de statistiques de l'Université du Washington, sous la direction de Matthew Stephens de 2004 à 2006. Mon arrivée dans ce laboratoire a correspondu à la publication par ce groupe d'un modèle d'approximation du coalescent (Li et Stephens, 2003) qui est encore aujourd'hui la base de développements méthodologiques très généralistes (e.g Delaneau *et al.*, 2012 ; Speidel *et al.*, 2019). A la même époque des programmes de caractérisation des polymorphismes des populations humaines furent initiés (*e.g.* le projet Human HapMap en 2003) suivis rapidement par les premières études sur la détection des effets de la sélection positive dans les populations humaines

(Sabeti *et al.*, 2007) ou la caractérisation des patrons de recombinaison humains (Crawford *et al.*, 2004) et de motifs d'ADN qui leurs sont associés (Myers *et al.*, 2005). Ce post-doctorat m'a permis d'apprendre énormément dans le domaine des statistiques, de la génétique des populations et des biostatistiques. Pendant que j'y développais des méthodes d'analyse d'association, Paul Scheet alors doctorant avec Matthew Stephens, travaillait de son côté sur un modèle de la diversité haplotypique (Scheet et Stephens, 2006) que j'utiliserais plus tard dans mon travail sur le développement de méthodes d'étude de la sélection. La fin de mon post-doctorat a correspondu aux premières publications d'études d'association pan-génomiques dans l'espèce humaine (The Wellcome Trust Case Control Consortium, 2007). A la fin de mon post-doctorat en Septembre 2006, j'ai été recruté au sein du département de génétique animale de l'INRA dans l'équipe de biomathématiques alors animée par Claude Chevalet. Mon arrivée à l'INRA correspondait au développement très rapide de la génomique chez les espèces animales, de mammifères en particulier. Les avancées technologiques développées dans le cadre des études de génétique humaine se sont vite déployées dans d'autres espèces. Ainsi le génome du bovin était publié en 2009, conjointement à une étude de la caractérisation de la diversité génétique des populations bovines (Zimin *et al.*, 2009 ; The Bovine HapMap Consortium, 2009). Les années suivantes virent la publication des génomes de toutes les espèces animales d'intérêt agronomique majeures et d'études de leur diversité génétique que je détaille en partie dans le cadre de la description de mes travaux de recherche dans la suite de ce mémoire. Ainsi, les avancées méthodologiques et technologiques développées dans le cadre de la génétique humaine ont largement contribué à faire progresser les recherches en génétique animale. A contrario, des modèles classiques développés dans le cadre de l'amélioration génétique des populations animales ont été introduits dans le domaine de l'étude des caractères quantitatifs humains (Yang *et al.*, 2010). Le développement de la génomique a donc eu pour conséquence que des communautés autrefois structurées par espèces aient vu leurs frontières s'abaisser. Il en est de même des frontières disciplinaires : les programmes de recherches actuels en génétique intègrent naturellement des questions de génomique, de génétique quantitative et de génétique évolutive. L'intégration de ces disciplines fait alors appel à la modélisation et en particulier à la modélisation statistique.

Dans ce mémoire je présente mes contributions à l'analyse statistique de données génétiques. Dans la suite de ce mémoire, mes publications scientifiques sont citées en **bleu** et celles qui sont issues des travaux des doctorants que j'ai contribué à encadrer sont de plus **sous-lignées**. J'ai structuré ce rapport par questions scientifiques plutôt que par déroulement temporel car il me semblait que cela offrait plus d'opportunités de montrer les relations entre mes différents travaux. Les questions scientifiques que j'ai abordé sont (i) la question de la cartographie des génomes, physique et génétique, (ii) l'étude des effets des mutations sur les caractères quantitatifs et (iii) l'étude de l'effet de la sélection phénotypique sur la diversité génétique des populations. Dans chacune de ces questions je décrirais les avancées méthodologiques auxquelles j'ai contribué ainsi que leur application sur des jeux de données réelles. Le dernier chapitre de ce mémoire présente les questions de recherche que je compte aborder à moyen terme.

1. Cartographie des génomes

1.1 Validation des assemblages de génomes par cartographie d'hybrides irradiés.

La séquence génomique d'une espèce est aujourd'hui une information fondamentale pour les études de génétique et de génomique. Elle fournit la liste exhaustive des locus présents dans le génome et leur organisation linéaire sur les chromosomes. L'établissement d'une séquence génomique complète, dite de référence, est donc une étape clé pour l'étude génétique d'une espèce. Jusqu'à très récemment, les techniques de séquençage haut-débit permettaient d'obtenir des informations parcellaires sur la séquence d'une espèce, en particulier car elles aboutissaient à la production de morceaux de séquence de petite taille (appelés scaffolds). L'établissement d'une séquence de référence assemblée en chromosomes entiers nécessite des informations supplémentaires permettant d'ordonner et d'orienter ces scaffolds. Chez les espèces animales, une méthode de cartographie physique, la cartographie d'hybrides irradiés (Radiated Hybrids : RH), a fourni une pièce du puzzle permettant d'obtenir des assemblages de meilleure qualité. Une partie de mon travail a consisté à développer des méthodes d'analyse pour améliorer l'information que fournissaient les cartes RH et participer à la validation des assemblages.

Dans une première section, je présenterai succinctement le principe de la cartographie RH et le modèle statistique existant sur lequel nous avons travaillé. Je présenterai ensuite le principe de ces méthodes et leurs avantages par rapport aux approches précédentes et finirai par montrer leur intérêt en prenant l'exemple de la validation de deux des trois assemblages de génomes sur lesquels j'ai pu travailler.

Cartographie RH intégrant des informations de génomique comparative.

Principe d'une expérience de cartographie RH Une expérience de cartographie par hybrides d'irradiation peut être décrite de la façon suivante. Tout d'abord des cellules de l'espèce étudiée sont irradiées ce qui conduit à la cassure physique des chromosomes en petits fragments. Ces fragments chromosomiques sont ensuite intégrés à des cellules vivantes d'une autre espèce receveuse (typiquement le hamster chez les animaux) où ils sont intégrés et peuvent survivre. Une cellule receveuse ayant intégré une partie des fragments obtenus (de l'ordre de 30%) est multipliée par mitose pour établir une lignée cellulaire appelée un **clone RH**. L'expérience aboutit finalement à un **panel RH** qui est un jeu de clones RH ayant chacun retenu une portion différente du génome.

La **cartographie RH** consiste ensuite à interroger par des techniques de génotypage la présence de certains marqueurs dans ce panel de clones pour estimer les liaisons physiques entre marqueurs. Il s'agit donc d'une méthode similaire à la cartographie génétique, les cassures physiques obtenues par irradiation jouant le même rôle que les cassures obtenues par recombinaison

dans des populations en ségrégation. Formellement après génotypage du panel RH, les données obtenues sont une matrice de présence (1) / absence (0) de tous les marqueurs dans tous les clones. La distance séparant deux marqueurs est définie à partir de leur co-occurence au sein des clones. En effet, pour une paire de marqueurs donnée quatre configurations sont possibles : (11),(10),(01) et (00). On peut caractériser les fréquences de ces configurations en fonction de deux paramètres : **le taux de rétention** r qui est la probabilité qu'un marqueur ait été retenu dans un clone ; **la fréquence de cassure** θ entre les deux marqueurs, analogue à un taux de recombinaison et qui fournit une mesure de la distance physique entre ces deux marqueurs. Sous certaines hypothèses, ces deux paramètres suffisent à modéliser les données et peuvent être estimés par maximum de vraisemblance indépendemment pour chaque paire de marqueurs.

Le but de la cartographie RH est d'établir un ordre entre marqueurs (une carte physique). A partir des analyses de chaque paire de marqueurs, il est possible de chercher l'ordre de vraisemblance maximum. Le problème n'est pas trivial car la combinatoire est importante, le nombre d'ordres possibles étant égal à $N!/2$ où N est le nombre de marqueurs. Il a pu être montré que le problème de recherche de l'ordre de vraisemblance maximum pouvait être formulé comme un problème du voyageur de commerce (Traveling Salesman Problem : TSP), et donc être résolu en utilisant des méthodes efficaces développées dans ce cadre.

Exploitation de l'information apportée par un génome proche pour la cartographie RH La structure des génomes évoluant lentement, les génomes d'espèces proches ont des organisations le long des chromosomes similaires. Ainsi deux gènes localisés à proximité l'un de l'autre chez une espèce ont tendance à l'être également chez une espèce proche. Lorsque c'est le cas, la synténie (présence des gènes sur le même chromosome) est dite conservée entre les deux espèces. Ce phénomène de conservation de synténie entre espèces peut être incorporé dans le modèle de cartographie RH pour définir une distribution a priori des ordres possibles. C'est le modèle Bayésien proposé par Faraut *et al.* (2007) :

$$P(\pi|X,\xi) \propto L(\pi;X,\xi)P(\pi|\pi_{ref}) \tag{1.1}$$

où X est la matrice des données RH, π est un ordre des marqueurs, ξ est le vecteur des paramètres du modèle (distances entre marqueurs θ et taux de rétention r) et π_{ref} est l'ordre a priori des marqueurs (*i.e.* chez une espèce proche où dans l'assemblage obtenu sans utiliser les données RH). Dans ce modèle $L(\pi;X,\xi)$ est la vraisemblance des données RH, calculée comme décrit succinctement ci-dessus, et $P(\pi|\pi_{ref})$ est une distribution *a priori* des ordres possibles, basée sur l'observation de l'ordre des marqueurs dans une espèce proche ou sur un assemblage n'ayant pas exploité les données RH.

Faraut *et al.* (2007) ont montré comment écrire la loi de distribution $P(\pi|\pi_{ref})$ et formulé le problème de recherche de l'ordre maximisant la probabilité $P(\pi|X,\xi)$ (et non plus uniquement la vraisemblance $L(\pi;X,\xi)$) comme un TSP, permettant sa résolution de façon numériquement efficace.

Estimation et exploitation de mesures d'incertitude des cartes RH

Le modèle de cartographie RH de Faraut *et al.* (2007) permet de construire efficacement une carte de plusieurs milliers de marqueurs, *i.e.* suffisamment résolutive pour pouvoir contribuer à l'amélioration des assemblages. Cependant, avec ce nombre de marqueurs, la meilleure carte obtenue a de grandes chances de ne pas correspondre au vrai ordre, en particulier lorsque les données RH ne sont pas complètes ou lorsqu'il y a des erreurs de génotypage. Nous avons pu montrer ce phénomène par simulations (Figure 1.1 tirée de **Servin *et al.* (2010)**). Lors de la comparaison de la carte obtenue avec un assemblage, les différences entre les deux ordres peuvent donc être dues soit à une erreur de carte soit à une erreur d'assemblage. En l'absence de mesure de confiance dans l'ordre RH, reconnaître les erreurs d'assemblage des erreurs de carte n'est pas possible.

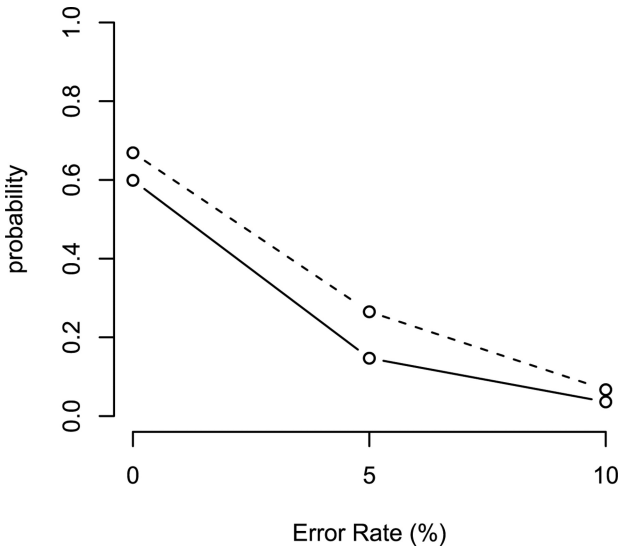


FIGURE 1.1 : Probabilité de retrouver la vraie carte en fonction du taux d'erreur simulé, en utilisant un modèle approché (ligne pleine) ou un modèle multipoint (ligne tiretée). *Simulations de 200 marqueurs dans 100 clones RH avec 10% de données manquantes*

L'idée de notre approche **(Servin *et al.*, 2010)** a été d'obtenir une distribution des ordres possibles puis d'exploiter cette distribution pour caractériser l'incertitude des cartes RH. Pour ce faire, nous avons développé un algorithme d'échantillonage MCMC des ordres possibles dans la distribution caractérisée par l'équation (1.1). A chaque étape de l'algorithme deux types de mise à jour de l'ordre courant sont utilisés : une mise à jour par Metropolis-Hastings basée sur l'inversion d'une partie des marqueurs de l'ordre courant et une mise à jour par échantillonage de Gibbs de la position des marqueurs. Cet algorithme est implémenté dans le logiciel de cartographie CarthaGène.

Cet algorithme MCMC produit une distribution d'ordres possibles pour les marqueurs d'un chromosome. Quand le nombre de marqueurs est grand, la distribution obtenue contient un grand nombre d'ordres possibles de probabilités comparables. Pour exploiter cette distribution, nous avons mis au point une méthode permettant d'extraire les propriétés d'ordre qui sont communes à l'ensemble des cartes de la distribution, d'en faire une représentation et de l'exploiter pour extraire ce que nous avons appelé une **carte robuste**. Le principe, illustré dans la figure 1.2, est de caractériser dans la distribution d'ordres, des groupes de marqueurs (i) dont l'ordre est conservé dans toutes les cartes, nous avons appelé ces groupes des *séquences* ou (ii) qui sont toujours regroupés ensemble mais dans des ordres différents, ces groupes sont appelés des *intervalles communs*. Une fois ces groupes caractérisés, la distribution d'ordres peut être représentée par un arbre d'inclusion, que nous avons appelé une *metamap*. A chaque noeud de l'arbre est associé un jeu de marqueurs ou de groupes de marqueurs et les probabilités de leurs ordres marginaux dans la distribution.

En partant le la metamap obtenue, il est ensuite possible d'extraire un sous-ensemble de marqueurs dont l'ordre est partagé par toutes les cartes échantillonnées ou plus généralement dont la probabilité de l'ordre a posteriori est supérieure à un seuil donné, ce que nous avons appelé une *carte robuste*. Par simulations, nous avons pu montrer que l'ordre des marqueurs contenus dans une telle carte robuste était quasi-exact, même en présence de taux d'erreurs de génotypage raisonnables. En revanche, le nombre de marqueurs présents dans une carte robuste diminue avec l'incertitude sur l'ordre. Par exemple quand les taux d'erreurs simulés étaient trop grands, nous avons montré que les cartes robustes obtenues contenaient très peu de marqueurs. Ainsi, une autre conclusion importante de notre étude était que la production de cartes robustes permet d'identifier les situations où l'inférence obtenue à partir des données RH est peu fiable.

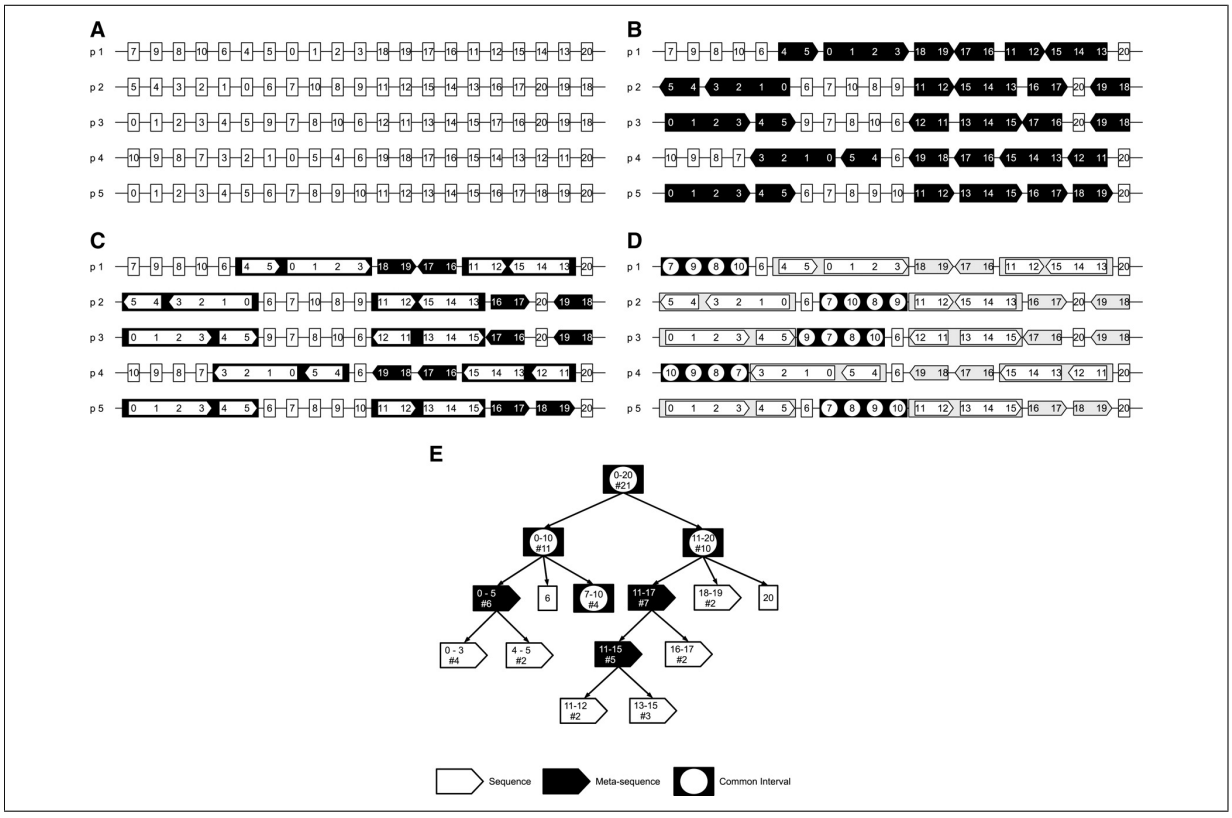


Figure 1.2 : Un exemple simple illustrant la construction d'une metamap (E) à partir d'une distribution d'ordres (A) par identification successive de séquences de marqueurs (B), de métaséquences (C) et d'intervalles communs (D). La metamap (E) est un arbre d'inclusion. L'incertitude dans l'ordre des marqueurs se trouve contenue dans les ordres possibles de chacun des descendants d'un noeud. Chaque noeud de la metamap est étiqueté par les marqueurs de début et de fin ainsi que le nombre de marqueurs qu'il contient.

Applications à trois espèces animales : le porc, le mouton et la chèvre

Amélioration de l'assemblage du génome du porc

Chez le porc, nous avons établi les premières cartes RH à partir d'une puce SNP d'environ 60 000 marqueurs **(Servin *et al.*, 2012)**. Les cartes finales contiennent près de 39 000 marqueurs. Elles ont permis de valider et de corriger la carte physique sur laquelle était basé le programme de séquençage du génome du porc. Nous avons pu corriger plusieurs erreurs dans cette carte, comme illustré par exemple dans la figure 1.3, et ainsi contribuer à l'amélioration de cet assemblage **(Groenen *et al.*, 2012)**.

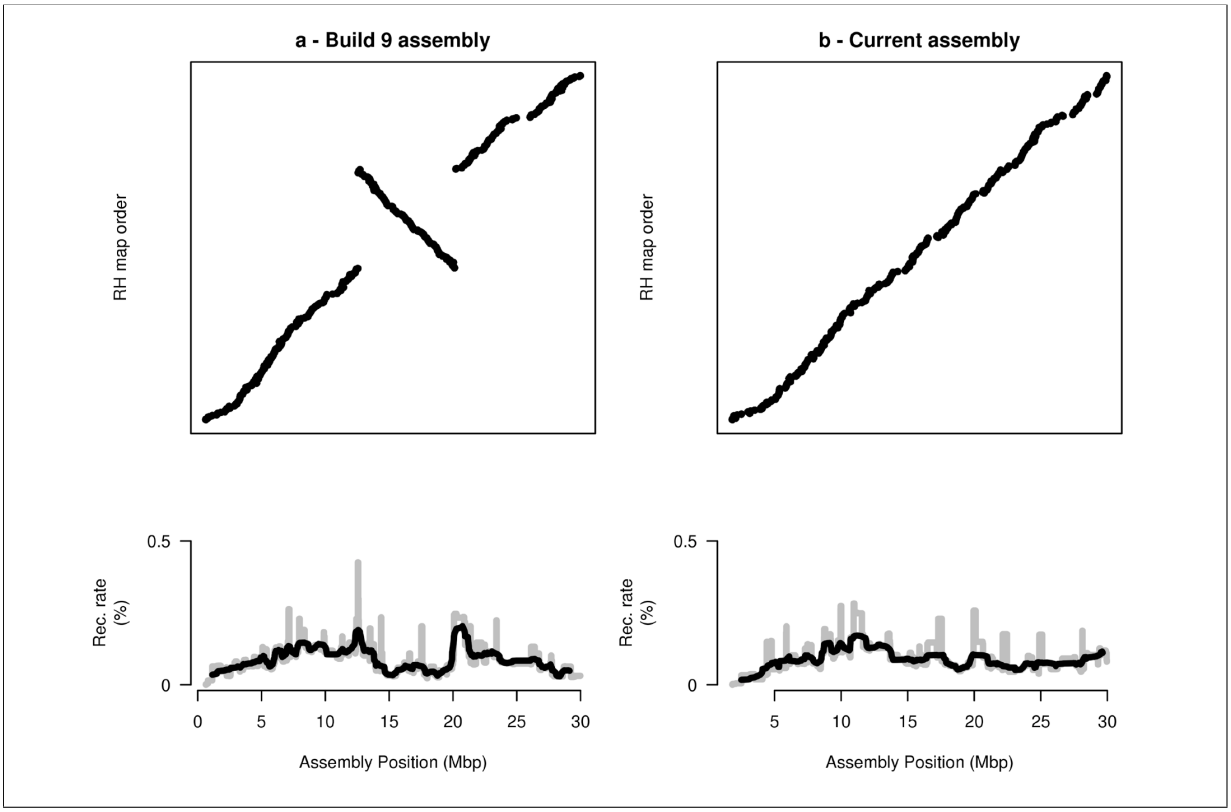
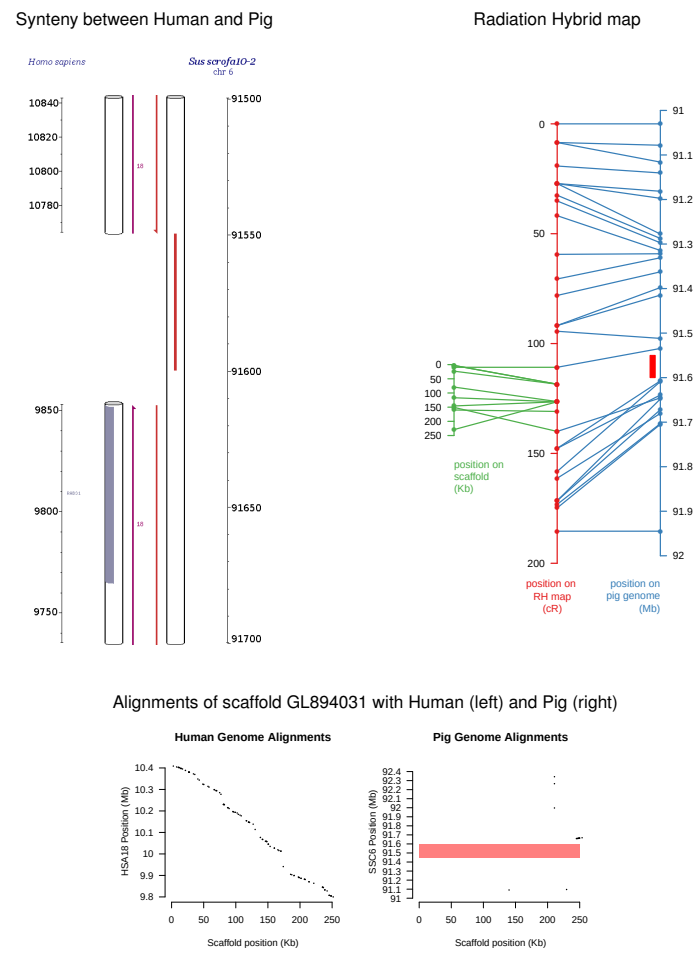


FIGURE 1.3 : Cette figure présente les 30 premières megabases du chromosome 6 porcin et sa comparaison avec la carte RH, pour l'assemblage version 9, à gauche, et l'assemblage de référence (version 10.2), à droite. Une estimation des taux de recombinaison sur ce segment sont représentés en bas. Cet exemple montre clairement que la résolution de l'inversion entre carte RH et assemblage à permis d'éliminer des pics de recombinaison aux frontières de l'inversion, validant la modification de la séquence de référence. Figure originale : Supplementary Figure 2 de Groenen *et al.* (2012).

Par ailleurs, nous avons utilisé ces cartes pour positionner sur le génome du porc un certain nombre de séquences (contigs, scaffolds) dont la localisation n'était pas connue. Un exemple est

présenté dans la figure 1.4. Au total l'ensemble de ces séquences représentait 72 megabases de séquence génomique supplémentaire dans l'assemblage.



Analysis of the predicted position of scaffold GL894031 on the pig genome RH maps position the 10 SNPs matching the scaffold between 91.1 and 91.9 Mb on SSC6. This localisation corresponds to a gap in the current pig genome (red rectangles on the figures). The syntenic human chromosome is HSA 18 (top left). The scaffold aligns nicely with the human chromosome but does not align to the current pig genome (bottom). A local RH map confirms linkage between the SNPs of the scaffold and the surrounding SNPs on the pig genome (top right), placing it in the gap.

Figure 1.4 : Exemple d'identification de la position génomique d'un scaffold du génome du porc en utilisant les cartes RH. Figure tirée de Servin *et al.* (2012).

Enfin, Flavie Tortereau, doctorante dans l'équipe de génétique porcine du LGC et chez Martien Groenen à l'université de Wageningen, a utilisé ces cartes RH et un logiciel d'estimation de taux de recombinaison que j'ai développée pour construire des cartes génétiques haute-densité du génome

porcin **(Tortereau *et al.*, 2012)** (cf. Section 1.2). Les cartes RH robustes ont été très utiles dans ce cas, car elles fournissaient un sous ensemble de marqueurs dont l'ordre pouvait être considéré comme fiable, et ainsi obtenir de bonnes estimations du taux de recombinaison, non biaisées par des erreurs d'assemblage.

Exploitation de plusieurs puces de génotypage pour l'assemblage du génome de la chèvre.

Chez la chèvre, des chercheurs de l'Université de Chine Centrale, Shuhong Zao et son doctorant Xiaoyong Du ont génotypé deux puces SNP d'espèces proches (la vache et le mouton) sur le panel RH disponible. Nous avons collaboré pour mettre au point les cartes RH de ces deux puces sur le génome de la chèvre, valider le premier assemblage de référence de cette espèce **(Dong *et al.*, 2013)** et *in fine* proposer une nouvelle version améliorée de cet assemblage **(Du *et al.*, 2014)**. Une des spécificités de cette étude a été la mise au point d'une technique spécifique pour déterminer les génotypes des clones RH à partir des données brutes (intensité de fluorescence) produites par la technologie utilisée. Cette procédure est illustrée dans la figure 1.5. Ces techniques d'imputa-

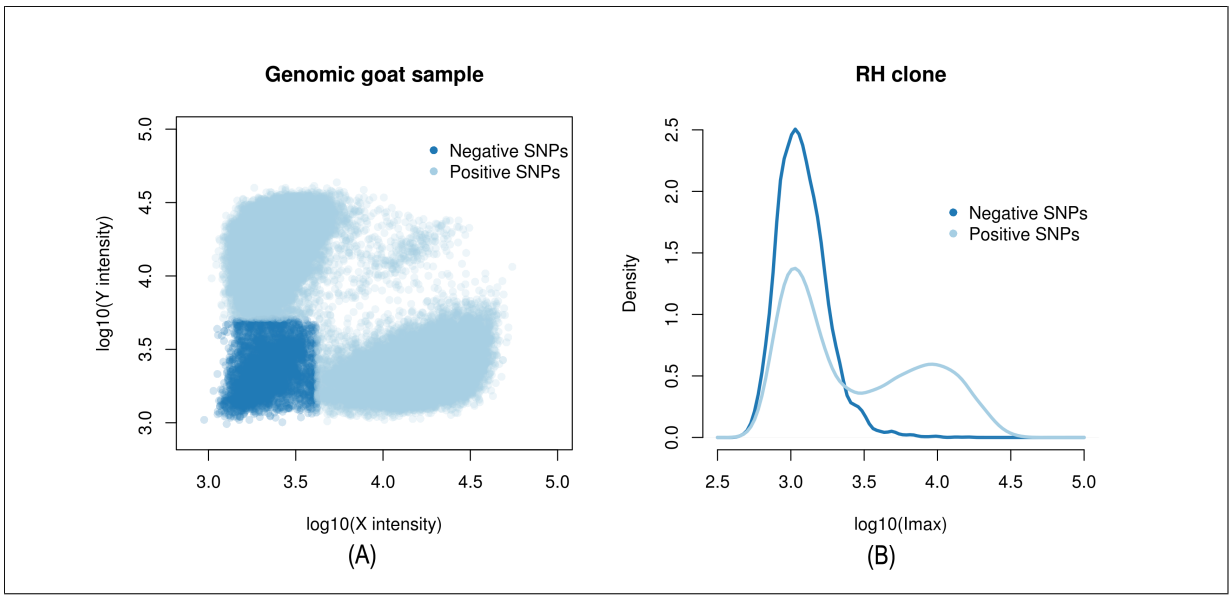


FIGURE 1.5 : Principe de l'imputation de génotypes dans le panel RH chèvre. A gauche : Un échantillon d'ADN génomique de chèvre est génotypé sur une puce d'une autre espèce (ici : le mouton). Les points représentent les intensités de fluorescence des deux allèles (un pour chaque axe) pour tous les marqueurs de la puce. Certains marqueurs (positive SNPs) de la puce (en bleu clair) fournissent un signal d'intensité fluorescente quand d'autres (en bleu foncé) non (negative SNPs). A droite : Distributions de l'intensité de fluorescence maximale entre les deux allèles. Lors du génotypage d'un clone RH les marqueurs négatifs sont utilisés pour définir une distribution d'intensité "background" à laquelle les autres marqueurs peuvent être comparés. La distribution des marqueurs positifs est un mélange de deux distributions, correspondant aux marqueurs retenus et non retenus dans le clone. Figure tirée de Du *et al.* (2014)

tion de génotype développées sur les données caprines m'ont permis de contribuer à la création de cartes RH du génome du mouton, exploitant l'information de deux panels RH différents (**Jiang *et al.*, 2014**).

Production d'assemblages de haute qualité avec les technologies de troisième génération

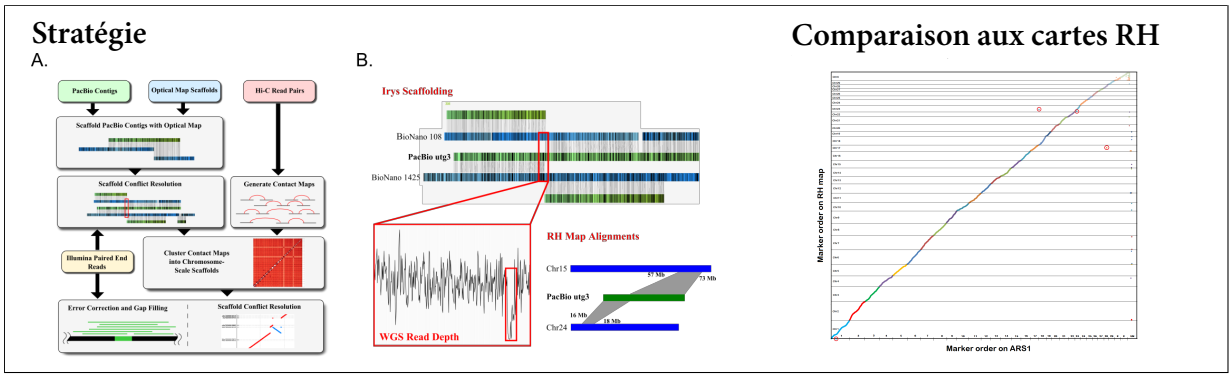


Figure 1.6 : Stratégie pour la création d'un nouvel assemblage du génome de la chèvre et comparaison avec les cartes RH. Figure tirée de Bickhart *et al.* (2017)

Aujourd'hui de nouvelles technologies de séquençage, dites de troisième génération, permettent de produire des lectures de très grande tailles (plusieurs kilobases). Ceci conduit à la possibilité de produire des séquences de référence (contigs) de très grande longueur uniquement sur la base des données de séquençage. Malgré cela, la production de séquences de chromosomes entiers nécessite toujours d'avoir recours à des expériences supplémentaires et à des informations de cartographie physique. Chez les espèces pour lesquelles nous avons produits des cartes RH, elles ont été utilisées pour valider ces nouveaux assemblages, comme illustré chez la chèvre sur la Figure 1.6 tirée de Bickhart *et al.* (2017). Cependant, pour les assemblages de nouvelles espèces, la cartographie RH est aujourd'hui une méthode de cartographie physique qui n'est plus suffisamment rapide et est trop onéreuse pour remplir ce rôle. Des technologies nouvelles, telle que la cartographie optique ou la capture de conformation chromosomique (Hi-C) sont venues la remplacer.

1.2 Cartes génétiques : modélisation et inférence de la recombinaison

La séquence génomique de référence d'une espèce est aujourd'hui un outil indispensable pour les études génétiques qui lui sont consacrées et plus généralement pour la compréhension de sa biologie. Cette séquence de référence fournit dans l'idéal une information complète sur l'ordonnancement des locus sur le génome et leurs distances *physiques*. En génétique, cette information est importante car lors de la transmission des chromosomes de parents vers leurs enfants, des locus proches vont avoir tendance à être transmis ensemble. Cependant, cette proximité ne se mesure

pas biologiquement en unités de distance physique (la paire de bases) mais en unité de distance génétique (le Morgan). La différence réside dans le fait que le mécanisme biologique qui casse les associations alléliques lors de la formation des gamètes, la *recombinaison*, n'est pas homogène le long du génome. Certaines régions génomiques recombinent plus que d'autres, c'est à dire que leurs probabilités d'être le lieu d'un échange chromosomique (un crossing-over), leurs *taux de recombinaison*, sont plus élevés. Au delà de la carte physique du génome d'une espèce, il est donc aussi important de produire des cartes génétiques, ou cartes de recombinaison, qui sont les cartes pertinentes pour les études génétiques.

L'établissement de cartes de recombinaison peut se faire avec différentes approches. Le taux de recombinaison au cours d'une méiose, dit taux de recombinaison *méiotique* (ici noté *c* et exprimé en centiMorgans par mégabase), peut être estimé en exploitant les patrons de ségrégations des allèles dans des familles. Chez les animaux d'élevage, la disponibilité récente de puces de génotypage de moyenne densité (de l'ordre de 50 à 100,000 marqueurs) a permis d'établir des cartes de recombinaison très résolutives, en particulier en exploitant les données générées par les programmes de sélection génomique. Si cette approche est largement utilisée, sa résolution est limitée par le nombre de méioses observables dans une population donnée et le nombre de marqueurs pouvant être génotypés, à la fois d'un point de vue technique mais également économique. L'estimation de taux de recombinaison à une échelle très fine (de l'ordre de la kilobases) nécessite donc d'utiliser d'autres approches. L'approche la plus couramment utilisée est basée sur le fait que les allèles de locus très proches portés par le même chromosome sont transmis en même temps lors d'une méiose. Plus ces locus sont proches (génétiquement), plus la recombinaison entre ces locus est rare, et plus ces associations entre allèles, *i.e.* leur *déséquilibre de liaison* (DL), vont perdurer longtemps. Des méthodes statistiques ont donc été développées qui exploitent le déséquilibre de liaison observé aujourd'hui pour en déduire les patrons de recombinaison le long du génome. Intrinsèquement, ces méthodes exploitent le résultat du processus de recombinaison méiotique cumulé sur de nombreuses générations et permettent donc d'établir des cartes de recombinaison très résolutives. Cependant, elles présentent certains inconvénients : d'une part elles ne fournissent pas d'estimation du taux de recombinaison méiotique mais d'un taux de recombinaison dit *historique* noté $\rho = 4Nc$, qui est aussi fonction de la taille efficace de la population (N), d'autre part, d'autres processus évolutifs que la recombinaison peuvent affecter les patrons de DL et donc l'estimation de taux de recombinaison historiques. C'est le cas en particulier de la sélection qui affecte ces patrons de manière hétérogène le long du génome, et l'interprétation des cartes de recombinaison historiques doit être faite avec prudence.

Dans cette section, je présenterai les travaux que j'ai effectués afin d'établir des cartes génétiques chez le porc et le mouton. Ces travaux ont été effectués dans le cadre de la thèse de Flavie Tortereau, au travers d'une collaboration avec l'université de Wageningen et dans le cadre de la thèse de Morgane Petit que je co-encadrais. Je montrerai en particulier l'intérêt de combiner les deux types d'approche mentionnés ci-dessus pour estimer des cartes de recombinaison fortement résolutives. Je reporterai au chapitre 2 (section 2.3) l'étude le la recombinaison en tant que processus biologique présentant une variation inter-individuelle.

Estimation des taux de recombinaison méiotiques

Le *taux de recombinaison* entre deux locus physiquement liés (*i.e.* sur le même chromosome) est la probabilité qu'un nombre impair de crossovers ait lieu au cours d'une méiose dans l'intervalle qui les sépare. Si ces marqueurs sont suffisamment proches, c'est essentiellement la probabilité qu'il y ait **un** crossover entre ces locus, et cette probabilité est alors égale à leur *distance génétique* exprimée en Morgans : une probabilité (= un taux) de recombinaison r de 0.01 correspond à une distance génétique de 0.01 Morgan, ou 1 *centiMorgan*. Lorsque l'on cherche à caractériser les variations d'intensité du processus de recombinaison le long d'un génome, les distances génétiques entre paires de locus successifs sont rapportées à leur distance physique. Ainsi, la mesure de l'intensité du processus de recombinaison est le taux de recombinaison méiotique (noté c) exprimé en centiMorgan par mégabase (cM/Mb). L'estimation de c consiste à (i) identifier des évènements de crossovers sur le génome et (ii) modéliser la distribution du nombre de ces crossovers sur des intervalles génomiques pour en estimer les taux de recombinaison.

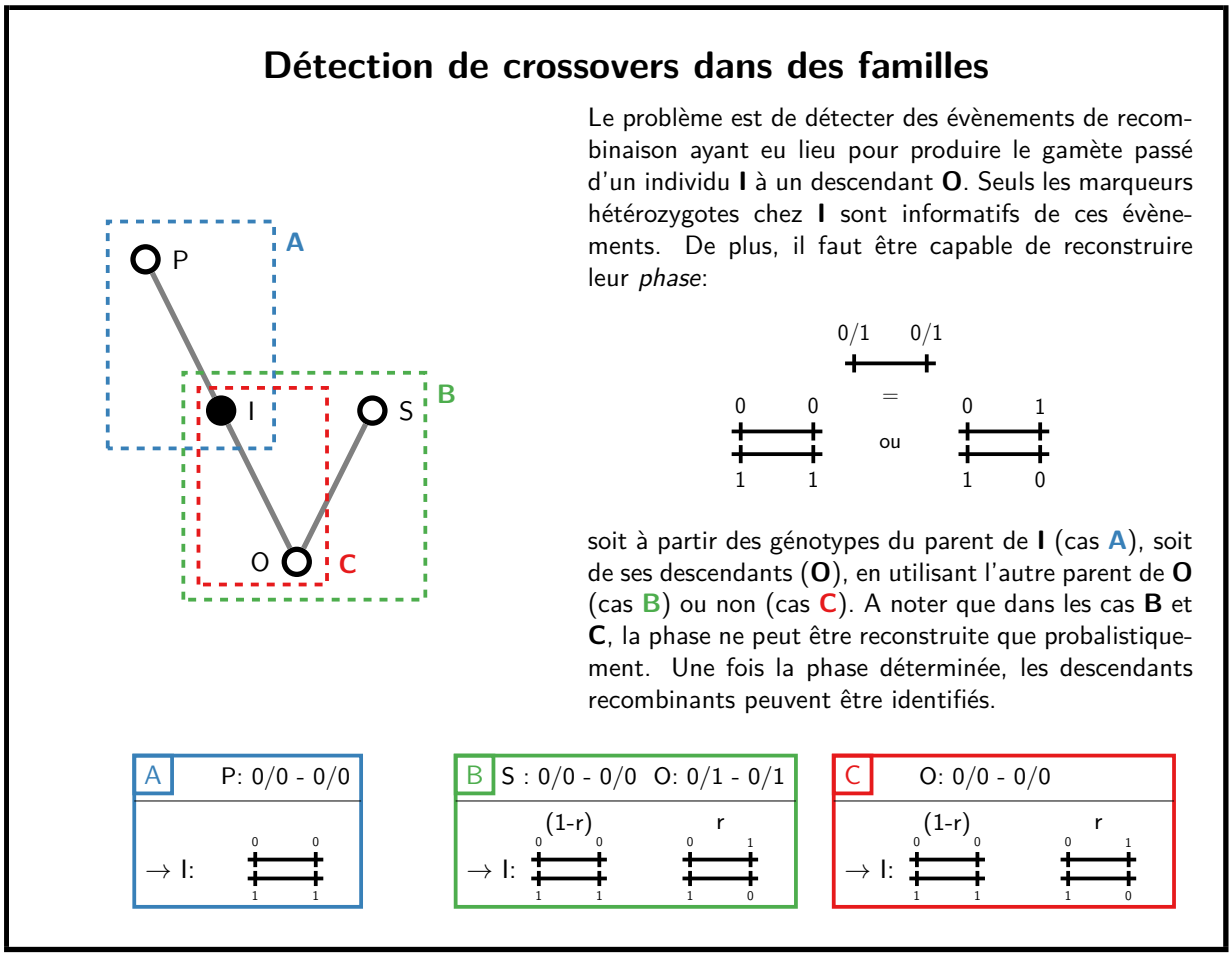


Figure 1.7 : Principe de détection des crossovers dans des familles

L'identification de crossovers le long du génome est obtenue par l'analyse des transmissions d'allèles dans des familles. Le principe en est illustré sur la figure 1.7. Ce principe a été utilisé pour caractériser la distribution des crossovers dans des familles humaines (Coop *et al.*, 2008) (cas B de la figure 1.7). Dans le cadre de notre étude de la recombinaison porcine **(Tortereau *et al.*, 2012)**, nous l'avons adapté pour l'analyse de familles de demi-frères (cas C de la figure 1.7). Dans ces approches, la reconstruction de la phase complète du parent I fait appel à des heuristiques pour combiner l'information obtenue pour chaque paire de marqueur en une phase globale d'un chromosome. Il est possible d'améliorer la reconstruction de la phase totale et donc la détection de crossovers en utilisant des algorithmes de satisfaction de contraintes pondérées (Favier, 2011) ou des modèles de Markov cachés (Druet et Georges, 2010; Fledel-Alon *et al.*, 2011; Druet et Georges, 2015).

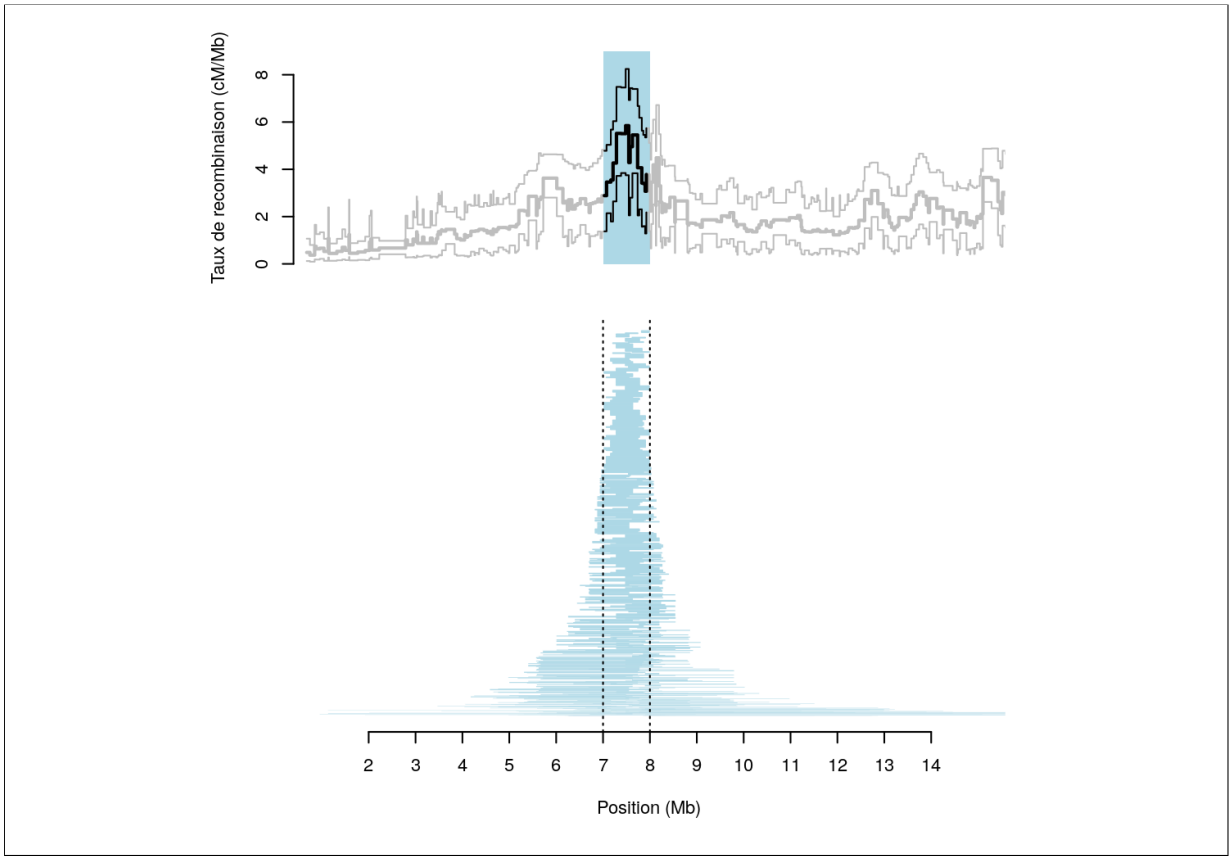


FIGURE 1.8 : **Illustration de l'estimation du taux de recombinaison à partir de crossovers chevauchants une région d'une mégabase.** Les crossovers chevauchant une fenêtre donnée (ici entre 7 et 8 megabase sur le chromosome 11 ovin) sont détectés à partir des données de génotypage (bas). L'ajustement d'un modèle statistique à ces crossovers permet ensuite d'estimer les taux de recombinaison dans la fenêtre (haut). Résultats tirés de Petit *et al.* (2017)

Modélisation de l'intensité de recombinaison et estimation du taux de recombinaison méiotique Ayant identifié des crossovers répartis sur le génome, nous avons proposé un modèle d'estimation du taux de recombinaison méiotique **(Petit *et al.*, 2017)** dans un petit [1] intervalle du génome (figure 1.8). Dans cet intervalle, la probabilité d'observer un crossover dans une méiose est de $0.01\,cl$, où c est exprimé en cM/Mb et l est la taille de l'intervalle en megabase. Dans un jeu de donnée constitué de M méioses, le nombre espéré de crossovers est de $0.01\,clM$. Enfin, notre modèle prend en compte le fait que l'intensité de crossovers varie entre individus, de telle sorte que pour un individu i, dont le nombre moyen de crossover par méiose est R_i, le nombre attendu de crossover dans l'intervalle est multiplié par le facteur R_i/R où R est le nombre moyen de crossovers parmi tous les individus du jeu de données. Une distribution naturelle pour modéliser le nombre de crossovers observé dans cet intervalle est la loi de Poisson :

$$y_i|c \sim \text{Poisson}(0.01\,clM_i\,R_i/R) \tag{1.2}$$

La vraisemblance totale pour le taux de recombinaison c, combinée sur l'ensemble des individus du jeu de données sera alors le produit des vraisemblances Poisson issues de l'équation (1.2). Pour compléter notre modélisation, nous combinons ensuite cette vraisemblance a une distribution a priori pour c :

$$c \sim \Gamma(\alpha, \beta) \tag{1.3}$$

A travers cet a priori, nous désirons exploiter l'information donnée par l'ensemble des intervalles du génome sur la distribution du taux de recombinaison. Ainsi, pour spécifier α et β, nous commençons par estimer de manière approché le taux de recombinaison c sur l'ensemble des intervalles, en utilisant la méthode de Sandor *et al.* (2012), puis nous ajustons ensuite une distribution Gamma sur les observations empiriques. Comme la loi Gamma est une distribution conjuguée de la loi de Poisson, la combinaison de l'a priori (1.3) avec les vraisemblances (1.2), implique que la distribution a posteriori de c est :

$$c|\, y_\bullet \sim \Gamma(\,\alpha\, + \sum_i y_i\,,\ \beta + \ 0.01\,l \sum_i M_i R_i/R) \tag{1.4}$$

Comme les méthodes de localisation des crossovers ne sont en général pas assez précises pour les affecter à un seul intervalle génomique (cf figure 1.8), pour finaliser notre modélisation nous avons ajouté une étape d'intégration de cette incertitude **(Petit *et al.*, 2017)**.

Modélisation du déséquilibre de liaison et estimation du taux de recombinaison historique

Bien que les tailles d'échantillon des dispositifs permettant d'estimer les taux de recombinaison méiotiques soient devenues très grandes, en particulier du fait de la mise en place de la sélection génomique dans de nombreuses espèces d'animaux d'élevage, la résolution des taux de recombinaison reste limitée à des intervalles de quelques dizaines voire centaines de kilobases. Améliorer

1. *petit* impliquant ici qu'il y a au plus un crossover intervenant dans l'intervalle au cours d'une méiose.

la précision d'estimation sur des intervalles plus petits nécessite d'étudier un nombre de méioses beaucoup plus grand que ne le permettent ces dispositifs. Pour ce faire, il faut se tourner vers d'autres approches. L'une d'entre elles consiste à étudier la recombinaison à partir de ses conséquences à long terme sur les associations d'allèles entre locus proches : le déséquilibre de liaison. N'ayant pas personnellement contribué aux développements méthodologiques dans ce domaine, je n'en présenterai ici que les principes et me contenterait de décrire succinctement le modèle de Li et Stephens (2003) que nous avons appliqué sur des données réelles (cf. Section 1.3).

Le modèle de Li et Stephens (2003) a été développé dans le but de fournir une approximation statistique calculatoirement efficace du processus de coalescence. Son application initiale était le phasage d'haplotypes d'individus diploïdes non apparentés et l'estimation de taux de recombinaison à une échelle fine. Pour simplifier, je décrirais ici la version haploïde du modèle en écartant la problèmatique de phasage quand les haplotypes ne sont pas observés. L'approximation de Li et Stephens (2003) est basée sur la décomposition de la vraisemblance d'une collection d'haplotypes $\mathcal{H}$:

$$P(\mathcal{H}|\theta,\rho) = P(h_1|\theta,\rho)P(h_2|h_1,\theta,\rho)\dots P(h_n|h_{n-1},\dots,h_2,h_1,\theta,\rho) \tag{1.5}$$

L'équation (1.5) est toujours vraie et explicite le principe de modélisation d'un nouvel haplotype comme conditionnée à des haplotypes connus. Elle dépend des paramètres populationnels de mutation ($\theta = 4N\mu$) et de recombinaison ($\rho = 4Nc$). Cependant, sous le modèle du coalescent de Kingman, ces probabilités conditionnelles ne sont pas connues et leur calcul est essentiellement impossible. Li et Stephens (2003) ont proposé de calculer $P(\mathcal{H}|\theta,\rho)$ par un produit de probabilités conditionnelles approximées. La vraisemblance obtenue est appellée vraisemblance PAC (pour Product of Approximation Conditionals) :

$$P(\mathcal{H}|\theta,\rho) \approx \pi(h_1|\theta,\rho)\pi(h_2|h_1,\theta,\rho)\dots\pi(h_n|h_{n-1},\dots,h_2,h_1,\theta,\rho) \tag{1.6}$$

La vraisemblance PAC (1.6) a pour principe de modéliser un nouvel haplotype h_n comme une mosaïque d'haplotypes connus $(h_1,\dots,h_{n-1})$. Ceci peut être illustré par la Figure 1.9 tirée de Li et Stephens (2003). Ces probabilités conditionnelles capturent les propriétés suivantes :

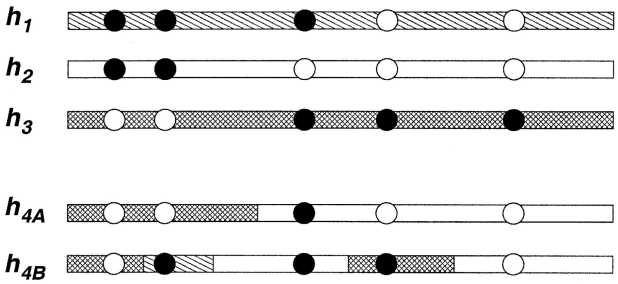


FIGURE 1.9 : Illustration du principe du modèle PAC tirée de Li et Stephens (2003). Un nouvel haplotype (h_{4A}) est modélisé comme un mosaïque d'haplotype connus (h_1, h_2, h_3).

— Un nouvel haplotype a plus de chance de ressembler à un haplotype observé fréquemment plutôt qu'à un haplotype rare

— La probabilité d'observer un haplotype nouveau décroît avec n
— La probabilité d'observer un haplotype nouveau augmente avec θ
— Si le nouvel haplotype n'est pas exactement le même q'un haplotype connu, il tend à en différer par un petit nombre de mutations plutôt qu'à en être complètement différent
— Du fait de la recombinaison, un nouvel haplotype ressemblera à des haplotypes connus sur des régions contiguës, ces régions étant d'autant plus longues que le taux de recombinaison local est faible.

D'un point de vue technique, le modèle PAC est un modèle de Markov caché (HMM) dont les états cachés sont la collection d'haplotypes "connus". Les probabilités d'émission dépendent du taux de mutation θ et les probabilités de transition du taux de recombinaison ρ. Pour calculer la vraisemblance d'un nouvel haplotype (1.6), les techniques classiques et calculatoirement efficaces propres aux modèles HMM comme l'algorithme Forward-Backward peuvent être utilisées.

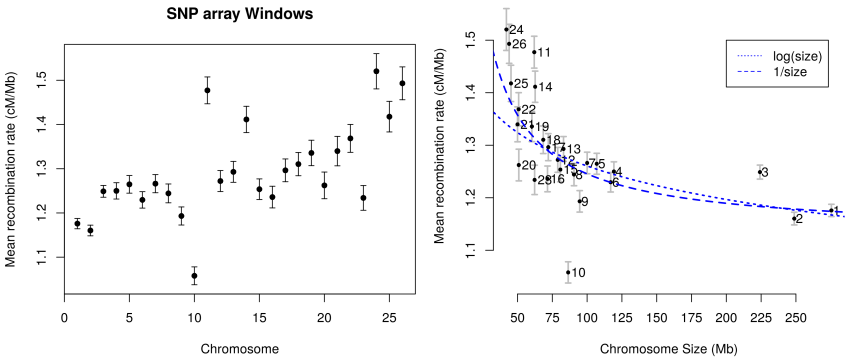
En se basant sur ce modèle Li et Stephens (2003) ont proposé un modèle Bayésien d'inférence du paramètre ρ, et plus particulièrement de sa variation le long d'une séquence génomique. Ce modèle est estimé en utilisant une approche MCMC (Markov Chain Monte Carlo) : pour chaque intervalle entre marqueurs l'estimation du modèle conduit à un échantillon de valeurs $\rho_j^{(k)} = \rho_w^{(k)} \lambda_j^{(k)}$ où ρ_w est le taux de recombinaison de base de la région et λ_j une intensité spécifique à l'intervalle, qui permet de capturer la forte hétérogénéité de la recombinaison à une échelle fine, due à la présence de points chauds de recombinaison.

1.3 Cartes fines de recombinaison chez le mouton

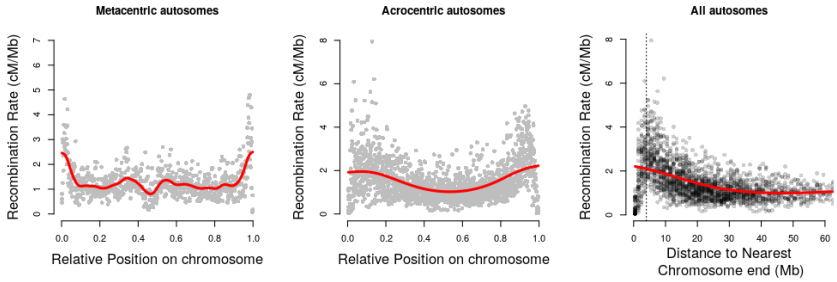
La mise en place de la sélection génomique, en particulier chez les ruminants, a conduit à la production de jeux de données très informatifs pour l'étude de la recombinaison. En effet les populations de référence sont constituées d'individus apparentés issus de pedigrees connus et génotypés pour des puces de génotypage de densité importante (un marqueur tous les 50 kilobases environ chez les mammifères). Dans le cadre de la thèse de Morgane Petit, nous avons exploité les données disponibles dans la population ovine Lacaune pour caractériser finement la distribution du taux de recombinaison le long du génome et cartographier les QTLs impliqués dans la variabilité individuelle tu taux de recombinaison (cf section 2.3). En plus de cette étude de la recombinaison méiotique, nous avons utilisé un jeu de données de génotypage haute densité **(Rochus *et al.*, 2018)** pour estimer les taux de recombinaison historiques dans cette même population. Finalement nous avons combiné ces deux inférences pour établir des cartes de recombinaison méiotique à forte résolution **(Petit *et al.*, 2017)**.

Cartes de recombinaison méiotique

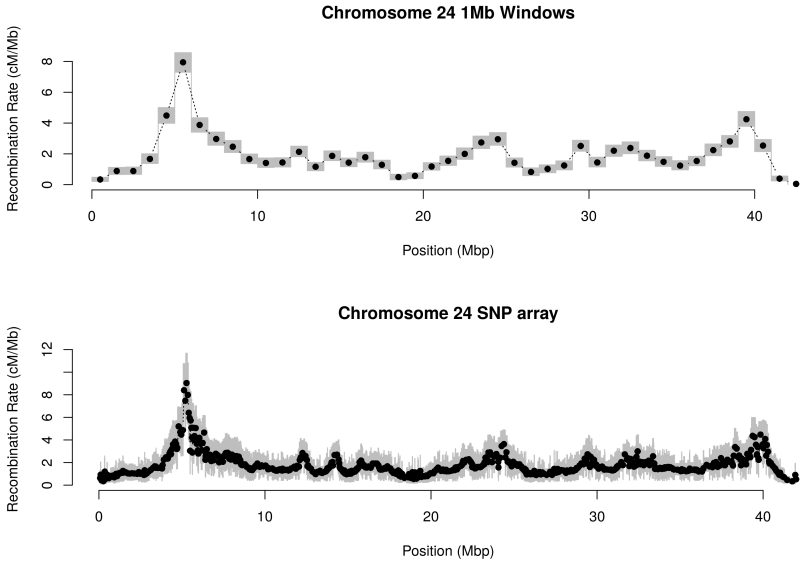
En utilisant la méthode décrite dans la section 1.2, nous avons établi des cartes de recombinaison du génome du mouton sur la base d'un jeu de données de 8000 individus mâles de la population Lacaune génotypés pour 54K marqueurs SNPs. Nous avons utilisés le logiciel LINKPHASE (Druet et Georges, 2015) pour identifier la localisation des crossing-overs dans ce pedigree. Nous avons ainsi identifié plus de 200.000 crossing overs répartis sur le génome, issus de 345 parents



(a) Taux de recombinaison moyens des chromosomes



(b) Taux de recombinaison le long des chromosomes

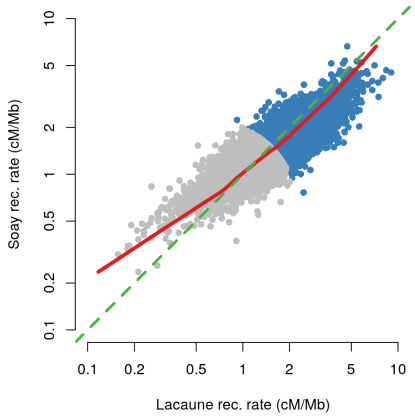


(c) Cartes de recombinaison méiotique du chromosome 24 ovin. haut : intervalles de 1 Mb, bas : intervalles des marqueurs de la puce SNP. Les estimations ponctuelles des taux de recombinaison sont en noir et les intervalles de crédibilité à 95% sont représentés en gris.

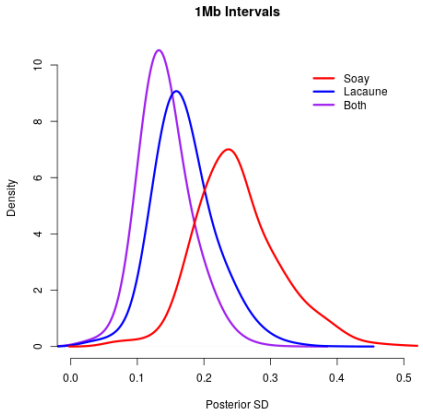
FIGURE 1.10 : Cartes de recombinaison méiotiques de la population ovine Lacaune.

mâles. En appliquant le modèle (1.4), nous avons établi les taux de recombinaison de chaque intervalle de la puce SNP utilisée (Figure 1.10c).

Nos résultats mettent en évidence une relation claire entre le taux de recombinaison moyen d'un chromosome et sa taille (Figure 1.10a), les petits chromosomes recombinant généralement plus que les grands. Ceci est cohérent avec l'hypothèse d'un crossing over obligatoire par bras chromosomique et par méiose. Nous pouvons cependant constater que certains chromosomes ont des comportement extrême comme par exemple le chromosome 10 qui a un taux de recombinaison beaucoup plus faible que sa taille ne pourrait le prédire. Ceci est due à une région génomique de grande taille au taux de recombinaison particulièrement faible sur ce chromosome. Du point de vue de la répartition des crossing-overs sur le chromosome, il y a une claire diminution du taux de recombinaisona avec la distance au télomère, et une chute de la recombinaison au niveau du centromère pour les chromosomes meta-centriques (Figure 1.10b).



(a) Taux de recombinaison de la population Soay (axe y) vs. ceux de la population Lacaune (axe x). La ligne verte est la droite $y = x$.



(b) Précision d'estimation des taux de recombinaison dans les jeux de données Lacaune, Soay et dans le jeu de données combinées. La ligne verte correspond à une égalité des variances à posteriori.

Figure 1.11 : Comparaison des cartes de recombinaison mâles dans les populations ovines Soay et Lacaune. Les points gris correspondent aux intervalles dont le taux de recombinaison moyen est < 1.5cM/Mb. Les lignes rouges sont les régressions locales (LOESS).

Alors que nous étudions le processus de recombinaison dans la population Lacaune, un autre groupe de chercheurs de l'université d'Edinbourg conduisait le même type d'étude dans une population très différente : la population Soay (Johnston *et al.*, 2016). Cette population vit sur une île des Hébrides extérieures en Écosse, l'île de Saint Kilda, et est une population férale qui fait l'objet d'études depuis de nombreuses années. Cette population est phylogénétiquement très éloignée de la population Lacaune. Les pedigrees y sont enregistrés et 3500 animaux (mâles et femelles) ont été génotypés sur la même puce SNP que celle utilisée pour notre étude. La publication de leurs résultats et de leurs données nous a offert l'opportunité de comparer les cartes de recombinaison dans ces deux populations. Pour ce faire, nous avons uniquement considéré les crossing-overs is-

sus de méioses mâles. Ce jeu de données était moins important que le notre, avec environ 88.000 crossing overs identifiés issus de méioses de 298 mâles. Nos analyses ont montré (Figure 1.11) que les deux populations avaient des taux de recombinaison mâles très comparables. Ainsi, la distribution et l'intensité de la recombinaison le long des chromosomes semble être conservée sur des échelles de temps relativement long (quelques milliers d'années ici) et malgré des histoires évolutives très différentes des populations. Ceci nous a conduit à combiner les deux jeux de données pour produire des cartes ayant une meilleure précision (Figure 1.11b).

Intégration avec les cartes de recombinaison historique

En plus de notre étude sur la recombinaison méiotique, nous avons estimé des taux de recombinaison historiques sur la base des patrons de déséquilibre de liaison dans un ensemble d'individus non apparentés de la même population Lacaune. En utilisant l'approche de Li et Stephens (2003) décrite ci-dessus nous avons estimé ces taux sur l'ensemble des intervalles de la puce ovine haute-densité. Environ 50.000 d'entre eux avaient des taux de recombinaison extrêmes comparés au reste du génome, suggérant qu'ils contenaient des points chauds de recombinaison. Ce nombre de points chauds est comparable à ce qui a été trouvé chez l'homme (de l'ordre de 25.000) bien qu'un peu plus élevé. Différentes hypothèses peuvent expliquer que notre estimation soit sur estimée **(Petit *et al.*, 2017)**.

Ayant établi deux types de cartes de recombinaison, nous avons cherché à les combiner pour profiter de la résolution apportée par les cartes historiques tout en corrigeant les effets démographiques qui les affectent. Nous aovons montré que la variabilité des taux de recombinaison méiotiques sur le génome était significativement associée aux différences de densité en points chaud. Ceci est illustré sur la Figure 1.12. Cette figure présente la comparaison des différentes cartes de recombinaison pour deux fenêtres du chromosome 24 très contrastées pour leur taux de recombinaison et pour leur densité en point chaud à courte échelle.

Pour aller plus loin, nous avons ensuite cherché à combiner statistiquement nos deux types d'inférence. Le principe est de considérer que les cartes méiotiques fournissent une estimation de c quand les cartes historiques fournissent une estimation de $\rho = 4N_e c$. Comme $log(\rho) = log(4N_e) + log(c)$, il est possible d'intégrer les deux types de cartes dans un modèle linéaire en y incluant un effet spécifique pour les estimations historiques qui capture le terme $log(4Ne)$ **(voir Petit *et al.*, 2017, pour les détails)**. Ce modèle nous a permis d'estimer la taille efficace de la population Lacaune à environ 7.000 diploïdes et nous avons estimé une corrélation de 0.73 entre les taux de recombinaison historiques et méiotiques, ce qui était une autre indication de la conservation des patrons de recombinaison au cours du temps. Enfin ce modèle nous a permis d'identifier dix régions du génome pour lesquelles les taux de recombinaison méiotiques et historiques étaient particulièrement éloignés. Pour la plupart de ces régions, ces différences tenaient d'effets des modèles statistiques utilisés, pour quatre d'entre elles ces différences semblent provenir d'effets sélectifs sur des gènes candidats souvent associés à des effets sélectifs pour des caractères de morphologie, de couleur de robe ou de déterminisme des cornes (cf Section 3.4).

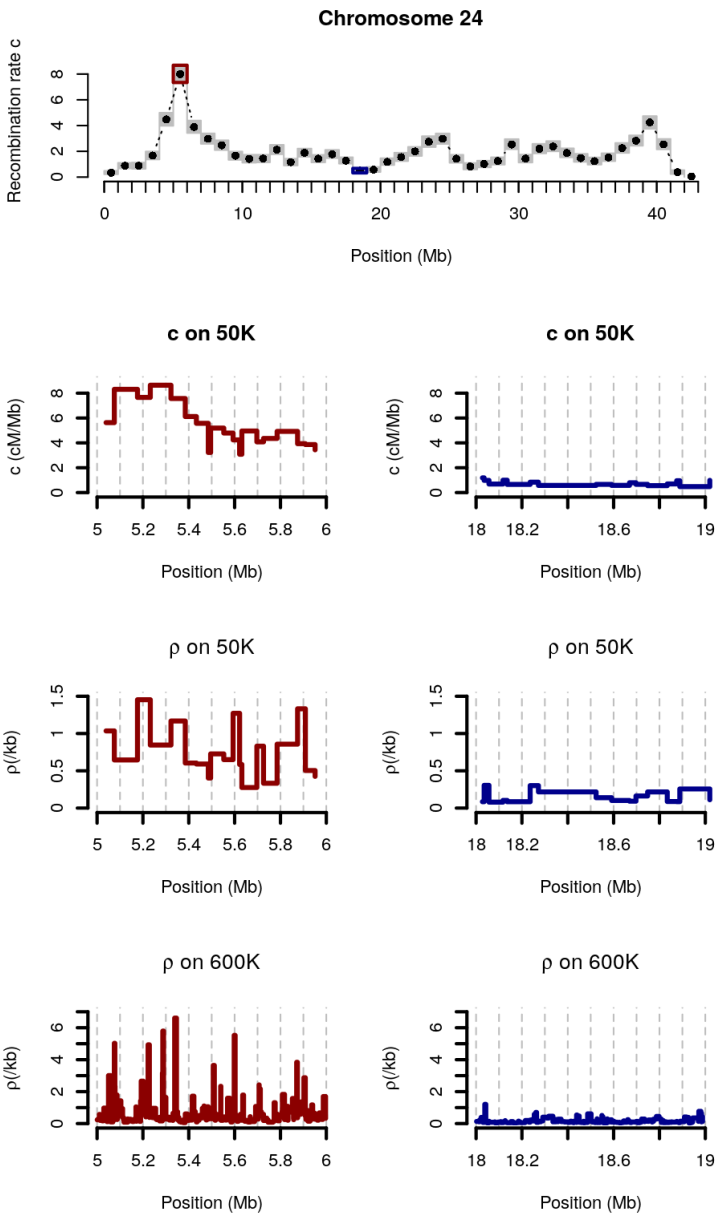


Figure 1.12 : Comparaison des taux de recombinaison historiques et métiotiques pour deux fenêtres d'1 Mb sur le chromosome 24 ovin. En haut : taux de recombinaison méiotique le long du chromosome. Deux fenêtres avec un fort (à gauche et en rouge) et un faible (à droite et en bleu) taux de recombinaison sont détaillées. Pour chaque panel, de bas en haut sont représentés : le taux de recombinaison méiotique des intervalles de la puce moyenne densité, les taux de recombinaison historiques pour ces mêmes intervalles et les taux de recombinaison historiques pour les intervalles de la puce haute-densité.

1.4 Exploitation en sélection

L'intérêt d'établir des cartes de recombinaison précises dans une espèce peut être illustré dans le cadre de l'optimisation de schémas de sélection destinés à produire des individus portant des combinaisons alléliques favorables : *la construction de génotype*. Dans le cadre de ma thèse j'ai travaillé à définir des méthodes permettant, à partir d'une carte génétique connue, de proposer des schémas de croisements permettant d'aboutir à un individu porteur d'un génotype idéal, appelé *idéotype*. Le principe se base sur la possibilité de calculer les probabilités d'obtenir un génotype multilocus quelconque à partir des parents, en utilisant les formules de récurrence de Hospital *et al.* (1996). Ces probabilités nécessitent de connaître la carte génétique de l'espèce.

Nous avons implémentés ces formules pour calculer les probabilités de génotypes multilocus dans des plans de croisements complexes à partir de lignées homozygotes et pour des espèces admettant l'auto-fécondation (*i.e.* la plupart des espèces végétales de grande culture) **(Servin *et al.*, 2002)**. Nous avons ensuite exploité cette méthode pour optimiser l'utilisation de marqueurs dans le cadre de programmes d'introgression de gènes par rétrocroisement, d'un point de vue théorique **(Servin et Hospital, 2002 ; Servin, 2005)** mais également dans le cadre de programmes expérimentaux **(Thabuis *et al.*, 2004 ; Lecomte *et al.*, 2004)**. Plus généralement, la connaissance des cartes génétiques d'une espèce est aussi fondamentale pour l'optimisation de programmes de pyramidage de gènes **(Servin *et al.*, 2004)**.

Un autre intérêt de l'établissement de cartes de recombinaison sur des puces de forte densité est de contribuer à définir les marqueurs à placer sur des puces de plus faible densité. Dans le cadre de nos travaux sur le mouton, nous avons ainsi montré l'intérêt de prendre en compte cette information pour optimiser la performance de l'imputation de génotypes **(Petit *et al.*, 2018a, cf. section 2.2)**.

1.5 Conclusion

Mes recherches sur l'établissement de cartes génomiques précises m'ont tout d'abord conduit à participer à l'amélioration des assemblages de plusieurs espèces animales. Ces assemblages procurent les cartes physiques détaillées des génomes et sont un outil majeur pour l'étude de la biologie d'une espèce. Aujourd'hui les technologies de séquençage de troisième génération permettent d'obtenir des assemblages de très haute-qualité pour un coût de plus en plus faible. Ceci ouvre la porte à de nombreuses opportunités en élargissant le panel d'espèces qu'il est possible d'étudier dans des approches de cartographie comparée. Avoir des assemblages de bonne qualité permet également d'établir des cartes de recombinaison fiables et très précises comme nous l'avons fait chez le mouton. Ces travaux ont montré que la distribution de la recombinaison était stable à l'échelle de quelques milliers d'années. Je compte aujourd'hui établir ce type de cartes chez d'autres espèces proches (chèvre, vache, ...) pour pouvoir étudier leur évolution sur des échelles de temps plus longues. L'objectif sera alors de mieux comprendre les contraintes qui affectent la répartition des crossing-overs sur le génome (cf. chapitre 4).

2. Du génotype au phénotype : déterminisme génétique des caractères complexes

2.1 Détection de QTL : analyse de liaison et analyse d'association

Une des questions majeures en génétique est de comprendre quelle part de la variation observée entre individus pour un caractère est due à leurs différences génétiques. Dans ce cadre, un des objectifs importants est d'identifier les mutations et les gènes ayant des effets sur ce caractère, *i.e.* la détection de QTL (Quantitative Trait Locus). D'un point de vue statistique, l'approche générale pour la détection de QTLs est de considérer qu'il existe une relation linéaire entre le génotype au QTL et le phénotype :

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2.1}$$

où $\boldsymbol{y}$ est le vecteur d'observation du phénotype μ est la moyenne générale du caractère, $\boldsymbol{X}$ est la matrice des génotypes au QTL, $\boldsymbol{\beta}$ est l'effet du QTL et $\boldsymbol{e}$ est la résiduelle [1]. Dans le cadre d'un caractère quantitatif, la distribution de la résiduelle est supposée Gaussienne $\mathcal{N}(0, \boldsymbol{R})$, conduisant à une vraisemblance également Gaussienne. Lorsque l'on cherche à estimer la localisation et l'effet d'un QTL (*i.e.* détecter un QTL), le principe est d'estimer le modèle (2.1) le long du génome et de tester s'il est meilleur que le modèle nul $\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{e}$. La formulation du modèle (2.1) n'est pas complète, en effet pour terminer sa spécification il convient de préciser (i) ce que l'on entend par génotypes au QTL (spécification de la matrice $\boldsymbol{X}$), (ii) comment l'on modélise ses effets $\boldsymbol{\beta}$ et (iii) la forme de la variance "résiduelle" $\boldsymbol{R}$. Ici il faut entendre *résiduelle* par rapport aux effets QTL. Ainsi, comme nous le verrons plus loin, cette variance peut inclure des effets génétiques s'ils ne sont pas dus au QTL. La forme finale du modèle statistique (2.1) va dépendre à la fois du type de dispositif expérimental considéré et de choix de modélisation. La littérature sur ce sujet étant très vaste, je ne présenterai dans cette section que les approches qui sont liées à mes travaux dans ce domaine.

Analyse de liaison

Le principe général de l'analyse de liaison est de considérer des pedigrees dans lesquels ségrègent des allèles issus des fondateurs et au sein desquels des phénotypes sont mesurés sur tout ou partie des individus. Ainsi, dans ce cadre, le génotype au QTL (*i.e.* une ligne de la matrice $\boldsymbol{X}$) indique qui sont les deux chromosomes fondateurs à l'origine des allèles portés par un individu.

1. Une formulation plus générale autoriserait d'autres effets fixes non génétiques mais n'a pas été intégrée ici par souci de clarté.

Croisements bi-parentaux Dans des plans de croisements bi-parentaux issus de deux lignées homozygotes (F2, rétrocroisement ...), le nombre d'allèles au QTL est de deux, elles sont généralement notées q et Q et trois génotypes sont possibles : qq, qQ et QQ. Pour ce type de croisement, les probabilités des différents génotypes au QTL à une position donnée du génome peuvent être calculées formellement à partir de la connaissance d'une carte génétique et des génotypes aux marqueurs observés. La relation exacte entre ces probabilités et la matrice $\boldsymbol{X}$ dépendra du modèle de déterminisme génétique considéré, comme j'en discuterai plus loin. Une fois cette matrice déterminée, des approches standards d'estimation et de test pour le modèle linéaire peuvent être utilisées (Lander et Botstein, 1989 ; Haley et Knott, 1992).

Cas de croisements multi-parentaux L'utilisation de croisements bi-parentaux pour la détection de QTL, si elle est simple et relativement puissante, est limitée pour plusieurs raisons. Tout d'abord l'utilisation de parents homozygotes n'est pas toujours possible, par exemple chez la majorité des espèce animales. D'autre part, même chez les espèces qui le permettent, la faible diversité génétique et phénotypique des croisements bi-parentaux peut limiter l'intérêt des QTLs détectés et l'on peut alors plutôt recourir à des pedigrees faisant intervenir plus de fondateurs **(Crepieux *et al.*, 2004)**(Churchill *et al.*, 2004).

Ainsi, une approche plus générale est requise pour des situations où le dispositif de détection de QTL (i) est constitué d'un pedigree complexe faisant intervenir plus de deux fondateurs et/ou (ii) où les fondateurs sont des individus hétérozygotes tous distincts. Dans ce dernier cas, comme les fondateurs sont chacun potentiellement porteurs de deux allèles différents au QTL il peut exister plus de deux allèles au QTL. D'un point de vue statistique ces dispositifs ont deux conséquences majeures qui aboutissent à utiliser des modèles linéaires mixtes (Almasy et Blangero, 1998 ; Xie *et al.*, 1998 ; George *et al.*, 2000).

Premièrement, dans des pedigrees complexes, les apparentements entre individus sur lesquels sont mesurés les phénotypes sont hétérogènes. Ainsi, les facteurs génétiques influençant le caractère et situés ailleurs qu'au niveau du QTL vont être responsable de covariances phénotypiques indépendantes des génotypes au QTL (Fisher, 1918). Dans le cadre du modèle (2.1) et en supposant des effets génétiques uniquement additifs, ceci conduit à modéliser la variance résiduelle

$$\boldsymbol{R} = \boldsymbol{A}\sigma_a^2 + \boldsymbol{I}\sigma_e^2 \tag{2.2}$$

où $\boldsymbol{A}$ est une matrice qui dépend des coefficients de parenté entre individus et des coefficients de consanguinité, σ_a^2 est la composante de variance génétique additive et σ_e^2 la composante de variance résiduelle *stricto sensu*.

Deuxièmement, du fait du grand nombre de fondateurs, dans le modèle (2.1) le vecteur $\boldsymbol{\beta}$ contient de nombreux effets qu'il peut être judicieux de modéliser à l'aide d'effets aléatoires (*i.e.* $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma_q^2)$) et la matrice de variance covariance des observations devient

$$V(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{X}^t\sigma_q^2 + \boldsymbol{R}$$

Ici, la matrice $\boldsymbol{X}\boldsymbol{X}^t$ peut être vue comme une matrice d'apparentement locale, dite d'*identité par descendance (IBD)*, au QTL. Ainsi le modèle (2.1) peut être réécrit en terme d'IBD :

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{g} + \boldsymbol{e} \tag{2.3}$$

avec $\boldsymbol{g} \sim N(\mathbf{0}, \boldsymbol{G}\sigma_q^2)$. Le problème d'estimation des génotypes au QTL devient alors de calculer la matrice $\boldsymbol{G}$ de probabilité d'identité par descendance au QTL.

Dans le cadre de croisements avec auto-fécondation, Xie *et al.* (1998) ont montré comment pouvait s'écrire la matrice $\boldsymbol{G}$ en fonction des probabilités des différents génotypes au QTL $\boldsymbol{p}$. Dans le cadre de la combinaison de multiples croisements bi-parentaux connectés, nous **(Crepieux *et al.*, 2004)** avons montré comment il était possible d'utiliser l'approche de Hospital *et al.* (1996) implémentée dans **Servin *et al.* (2002)** (cf. section 1.4) pour calculer ces probabilités et ainsi permettre d'y détecter des QTLs. Notre travail fournissait donc les outils permettant d'appliquer la détection de QTLs dans des dispositifs existants issus de programmes de sélection classique.

Analyse d'association

L'analyse de liaison a permis d'identifier un grand nombre de gènes majeurs et de QTL d'effets forts mais à la fin des années 90, un constat commençait à apparaître concernant la puissance limitée de cette approche. Ainsi, dans un article visionnaire Risch et Merikangas (1996) démontraient qu'une approche basée sur l'*association* génétique offrait une puissance bien supérieure à l'analyse de liaison et était probablement requise pour détecter des QTL d'effets faibles. La génétique d'association a pour principe de non plus tester la relation entre les ségrégations des segments chromosomiques dans des familles mais de tester directement la relation entre le génotype observé à un polymorphisme et le phénotype. L'hypothèse sous-jacente à ce test est que le polymorphisme testé est une mutation biologiquement causale, ou tout le moins que les génotypes à ce polymorphisme sont corrélés à ceux d'une mutation causale (*i.e.* les deux polymorphismes sont en *déséquilibre de liaison*, abb. DL).

Le plan d'expérience idéal d'une étude d'association consiste donc à échantillonner des individus non apparentés, et à les génotyper pour un nombre suffisant de polymorphismes de telle sorte que chaque QTL soit en DL avec au moins l'un d'entre eux. Comme le déséquilibre de liaison entre polymorphismes dans une population décroît rapidement avec la distance génétique, il est nécessaire de pouvoir génotyper de très nombreux polymorphismes pour que cette approche puisse fonctionner. Ainsi, Risch et Merikangas (1996) notaient que la principale limitation de cette approche n'était pas dans la difficulté d'échantillonage ou dans sa puissance statistique mais dans les capacités technologiques de l'époque pour pouvoir génotyper tous les QTL d'un génome. Cet article concluait de la façon suivante : "*C'est à la technologie en biologie moléculaire de développer les outils permettant de relever ce défi et d'amener l'information nécessaire à l'identification du déterminisme génétique des maladies humaines complexes.*" [2]

Aujourd'hui, il est clair que ce défi technologique a été relevé avec le développement de puces de génotypage et de séquenceurs haut-débit. Ces technologies ont été très rapidement utilisées pour caractériser la diversité génétique globale des populations humaines d'abord puis de nombreuses espèces ensuite. La mise a disposition de ces données et leur utilisation dans des échantillons de grande taille a bien évidemment aboutit au développement de nombreuses méthodes

2. traduction personnelle

statistiques pour modéliser l'association génotype / phénotype à partir de ces données. Dans le cadre de mon post-doctorat, j'ai travaillé à développer ce type de méthode et en particulier à la mise en place de méthodes dites d'association par *imputation de génotypes*.

2.2 Méthode d'analyse d'association par imputation de génotypes

Le développement de puces de génotypage haute-densité a permis de mettre en oeuvre les études d'association pan-génomiques. Cependant, les premières analyses de ce type ne testaient qu'un sous-ensemble des polymorphismes existants dans une population : ceux présents sur la puce utilisée. Du fait du déséquilibre de liaison, tester ces polymorphismes permettaient d'assurer une certaine puissance de détection. Cependant, idéalement il aurait fallu pouvoir tester directement l'ensemble des polymorphismes existants.

Dans le cadre de mon travail de post-doctorat, nous avons proposé l'idée de profiter des résultats d'études génériques de la diversité d'une population, de type HapMap (International Hapmap Consortium, 2007) ou 1000 génomes (The 1000 Genomes Project Consortium, 2015), qui permettent d'y décrire précisément la diversité génétique et la structure du DL, pour prédire les génotypes des individus échantillonés dans le cadre d'une étude d'association. Ce processus de prédiction est appelé *imputation de génotypes*. Plus précisément, nous avons montré comment utiliser des modèles statistiques du déséquilibre de liaison pour imputer les génotypes et comment utiliser les résultats de cette imputation dans un modèle de régression Bayésienne de l'association génotype / phénotype **(Servin et Stephens, 2007)**.

Dans cette section, je présenterai dans un premier temps notre modèle de régression Bayésienne simple marqueur, puis son extension pour analyser conjointement plusieurs marqueurs et enfin l'incorporation de l'imputation de génotypes. Je conclurai sur l'état actuel des méthodes statistiques d'analyse d'association.

Régression Bayésienne simple marqueur

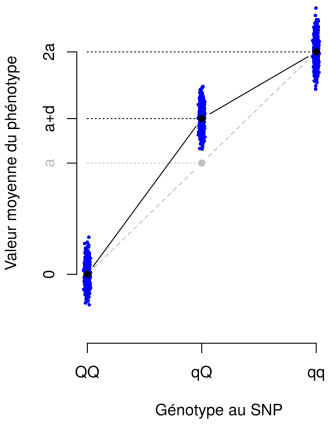
Nous considérons un ensemble de n individus pour lequel nous disposons de l'information sur un phénotype quantitatif $\boldsymbol{y} = (y_1, \ldots, y_n)$, et de leurs génotypes à un ensemble de n_s SNPs (matrice $\boldsymbol{G}$ de dimension $n \times n_s$). Nous modélisons l'association avec le génotype à un marqueur bi-allélique (Single Nucleotide Polymorphism, abb. SNP) par une régression linéaire (2.1) :

$$y_i = \mu + \boldsymbol{x_i}\boldsymbol{\beta} + e_i \tag{2.4}$$

où μ est la moyenne du génotype des individus portant un génotype de référence, $\boldsymbol{x_i}$ est la ligne de la matrice d'incidence qui dépend du génotype de l'individu au marqueur (vecteur ligne $\boldsymbol{G_i}$) (voir ci-dessous), $\boldsymbol{\beta}$ les effets correspondant et e_i est la résiduelle. Nous faisons l'hypothèse que les e_i sont iid $\sim \mathcal{N}(0, 1/\tau)$, où τ est l'inverse de la variance, aussi appelée *précision*. Ainsi, dans notre modèle $y_i|\boldsymbol{x_i}, \mu, \boldsymbol{\beta}, \tau \sim \mathcal{N}(\mu + \boldsymbol{x_i}\boldsymbol{\beta}, 1/\tau)$.

Nous modélisons les effets d'un SNP en autorisant la dominance, les trois génotypes possibles (homozygote Q/Q, hétérozygote Q/q, homozygote q/q, où Q est l'allèle de référence) ayant pour

effets respectifs 0, $a + ak$, $2a$ (Figure 2.1). Ainsi la matrice $\boldsymbol{X}$ contient deux colonnes, l'une indiquant le nombre de copie de l'allèle Q et l'autre valant 1 pour les lignes correspondant à des individus hétérozygotes. Ainsi à un SNP est associé deux effets, un effet additif a et un effet de dominance $d = ak$.



Genotype	x_{i1}	x_{i2}	Effet
Q/Q	0	0	0
Q/q	1	1	$\beta_1 + \beta_2$
q/q	2	0	$2\beta_1$

FIGURE 2.1 : Modèle d'association d'un SNP avec le phénotype

Distributions *a priori* sur l'effet d'un SNP Nous avons choisi de traiter ce modèle en utilisant une approche Bayésienne qui permet d'intégrer à la modélisation une information spécifique au problème posé (ici l'association génotype / phénotype). Dans notre cas cette information portera entre autres sur l'amplitude attendue de l'effet d'un QTL. Une grande partie du processus de modélisation Bayésienne consiste donc à déterminer comment exprimer explicitement ces informations sous la forme de distributions dites *a priori* (*i.e.* ne dépendant pas des données). Dans le cadre du modèle (2.4), nous avons défini ces distributions en considérant que : (i) les résultats ne doivent pas dépendre de l'unité de mesure du phénotype (ii) s'il est désirable d'autoriser des déviations à l'additivité, celle-ci doit rester une hypothèse parcimonieuse favorisée et (iii) l'estimation du modèle doit rester calculatoirement efficace. Nous avons proposé deux types d'*a priori* permettant chacun de satisfaire en partie ces considérations. Le premier (nommé D1) spécifie un modèle de dominance plus souhaitable (cas (ii) ci-dessus) au détriment d'un coût calculatoire plus important, le second (nommé D2) fait une hypothèse simplificatrice sur les effets d'un SNP mais permet des calculs rapides.

Les *a priori* D1 et D2 partagent les mêmes distributions pour μ, τ et a. Pour obtenir une invariance des résultats d'analyse à une translation ou une multiplication du phénotype nous utilisons les *a priori* impropres de Jeffreys ($p(\mu, \tau) \propto 1/\tau$). La distribution *a priori* sur a est une loi normale $\mathcal{N}(0, \sigma_a^2/\tau)$, où σ_a reflète l'amplitude typique d'un QTL, mesuré en unité d'écart-type résiduel. Par exemple, spécifier $\sigma_a = 0.5$ implique qu'*a priori* un QTL a ≈95% de chance d'avoir un effet inférieur à 1 écart-type résiduel en valeur absolue. La spécification de σ_a va dépendre du

contexte. Dans des approches plus récentes d'analyse d'association pan-génomique, ce paramètre n'est d'ailleurs pas fixé mais estimé à partir des données (voir ci-dessous).

La différence entre les *a priori* D1 et D2 se trouve dans la modélisation de l'effet de dominance (Figure 2.2).

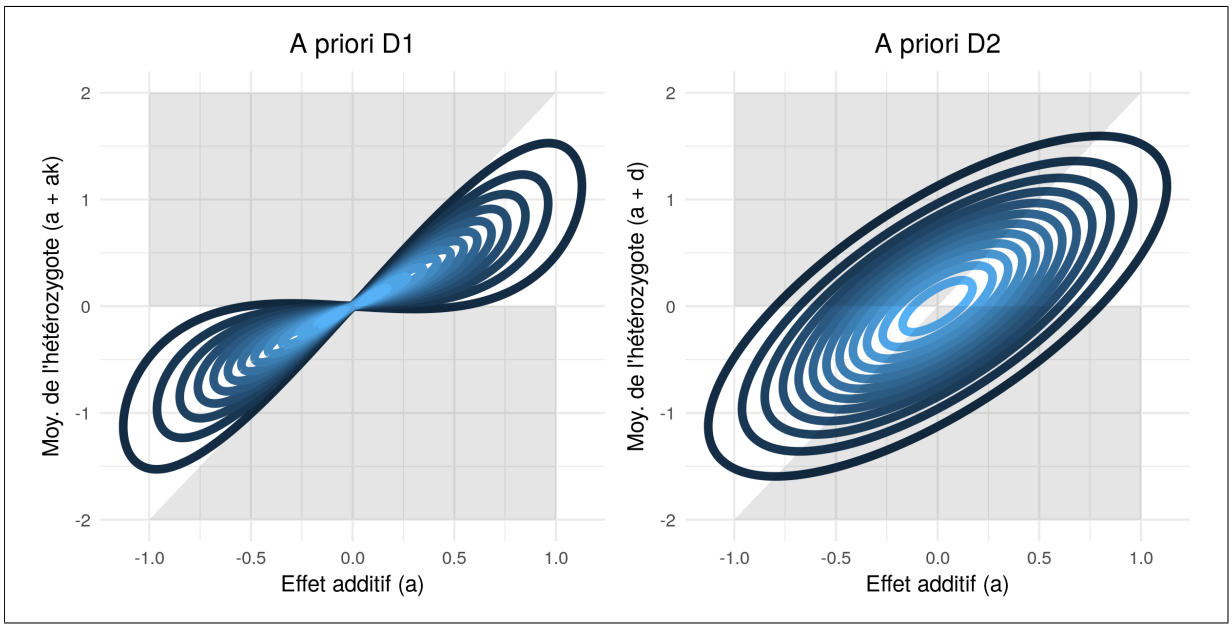


Figure 2.2 : **Deux différents *a priori* sur l'effet de dominance d'un SNP.** Densité conjointe de l'effet additif et de la moyenne d'un individu hétérozygote selon l'*a priori* D1 (gauche) ou D2 (droite). Les parties grisées correspondent aux zones de super-dominance (|a+d| > |2a|). L'*a priori* D1 contrôle leur probabilité *a priori* explicitement. L'*a priori* D2 fait l'hypothèse d'indépendance entre effet additif et effet de dominance, autorisant largement la super-dominance.

L'*a priori* D1 fait l'hypothèse que la super-dominance est rare ce qui semble biologiquement raisonable. Spécifiquement, l'effet de dominance d'un SNP (d ou β_2 dans la figure 2.1) a une forte probabilité d'être inférieur à son effet additif (a ou β_1 dans la figure 2.1).Ceci est obtenu en introduisant une dépendance entre a et d :$P(a,d) = P(d|a)P(a)$ avec $P(d = ak|a) \sim \mathcal{N}(0, a^2\sigma_k^2)$. Cette structure particulière a pour conséquence que l'estimation du modèle d'association nécessite de recourir à un algorithme d'estimation stochastique de type MCMC (cf Fichier additionnel 2 dans **Servin et Stephens (2007)**), relativement coûteux en temps de calcul .

L'*a priori* D2 a été conçu comme une approximation de D1, éliminant la dépendance entre a et d, de telle sorte que $P(a,d) = P(d)P(a)$. Ceci a pour conséquence de rendre analytiques et rapides les calculs de différentes statistiques (cf Fichier additionnel 1 dans **Servin et Stephens (2007)**). En revanche la structure des effets *a priori*, illustrés dans la figure 2.2, n'est pas idéale : une forte probabilité *a priori* est autorisée pour des cas de super-dominance et celle-ci est d'autant plus rare que les effets additifs sont forts.

Régression Bayésienne multiple

Dans notre article (Servin et Stephens, 2007), nous avons présenté le principe de notre modélisation dans le cadre de l'étude d'une région candidate, bien que les méthodes et le principe de l'imputation soit applicable à une analyse tout génome. Dans le cadre d'une étude locale, nous avons proposé un modèle qui autorise la présence de plusieurs QTL dans une même région génomique dont les effets s'additionnent (pas d'épistasie locale). Dans ce cas, le modèle (2.4) peut être étendu en une régression multiple :

$$y_i = \mu + \sum_j \boldsymbol{x}_{ij}\boldsymbol{\beta}_j + e_i \tag{2.5}$$

Ici, la somme indicée par j est effectuée sur un ensemble de SNP ayant un effet sur le phénotype. Les *a priori* sur les effets $\boldsymbol{\beta_j}$ sont ceux présentés ci-dessus.

A priori sur le nombre de SNP associés Dans le cadre de cette régression multiple, nous avons fait l'hypothèse que seuls certains SNP ont un effet non nul. Spécifiquement, nous avons supposé qu'avec une probabilité p_0 aucun SNP n'est associé au QTL et qu'avec une probabilité $(1 - p_0)$, l QTLs sont associés, l ayant une distribution sur un ensemble $1, 2, \ldots, n_s$ où n_s est le nombre total de SNPs. Cette distribution *a priori* sur le nombre de SNP causaux dans la région nécessite de spécifier de nouveaux paramètres : p_0 et la distribution $P(l)$ (des hyper-priors). Cette spécification peut dépendre du contexte. De plus, dans le cas d'une analyse pan-génomique p_0 peut être estimée à partir des données (voir ci-dessous).

Inférence

Ayant modélisé les effets QTLs dans une région génomique, l'inférence va porter sur deux questions. Premièrement la *détection* d'une association entre génotypes et phénotypes et deuxièmement, *l'interprétation* des associations observées. Notre modélisation permet de répondre conjointement à ces deux questions.

Détection d'une association En ce qui concerne la détection des associations, nous avons proposé d'utiliser le facteur de Bayes, abb. BF (Kass et Raftery, 1995), comme statistique pertinente :

$$BF = P(\boldsymbol{y}|\boldsymbol{G}, H_1)/P(\boldsymbol{y}|\boldsymbol{G}, H_0) \tag{2.6}$$

où H_0 dénote l'hypothèse nulle qu'aucun SNP n'est associé au phénotype et H_1 l'hypothèse complémentaire (au moins un SNP est associé au phénotype). Pour calculer cette statistique, il est nécessaire d'intégrer les paramètres inconnus. Dans le cas du prior D1, ceci peut se faire en utilisant les échantillons MCMC. Dans le cas du prior D2, le BF peut être calculé analytiquement. Dans le cas particulier où le modèle alternatif n'autorise au plus qu'un seul SNP associé ($P(l > 0) = P(l = 1) = (1 - p_0)$), l'équation (2.6) devient :

$$BF = (1/n_s) \sum_{j=1}^{n_s} P(\boldsymbol{y}|\boldsymbol{G}, H_j)/P(\boldsymbol{y}|\boldsymbol{G}, H_0) \tag{2.7}$$

où H_j correspond au cas où le SNP j est associé au phénotype. Ceci démontre un des intérêt de l'utilisation des BF : la combinaison du signal à plusieurs SNP est naturelle, contrairement à une inférence basée sur des p-valeurs.

D'un point de vue Bayésien, le BF est *la* mesure permettant d'évaluer la pertinence d'une hypothèse, si l'on accepte les choix des distributions *a priori*. Une façon de réconcilier les approches Bayésienne et fréquentiste est de calculer, par permutations, une p-valeur empirique associée au Bayes Factor. Cette approche est appelé le compromis Bayes / non-Bayes (Good, 1992).

Interprétation des associations Une fois le modèle (2.5) estimé sur les données, les associations obtenues peuvent être interprétées d'une part à la lumière des distributions *a posteriori* des effets additifs et dominants des SNPs mais également à la lumière des distributions des probabilités d'inclusion des SNPs dans le modèle. En effet, c'est un des avantages de la régression Bayésienne que de permettre d'estimer la probabilité qu'un SNP soit "causal". Spécifiquement, supposons qu'une région du génome contienne de nombreux SNPs aux génotypes très corrélés (en fort DL), chacun étant également associé fortement au phénotype, alors la conclusion correcte est que chacun des SNPs est potentiellement causal, sans qu'il soit possible d'identifier lequel. Dans le cadre de notre modèle, ceci conduire à une probabilité globale d'association très forte (*i.e.* un BF (2.6) très élevé) qui sera répartie équitablement entre tous les SNPs potentiels. Ceci permet par exemple d'établir une priorisation des candidats à la validation fonctionnelle.

Combinaison avec l'imputation de génotypes

Jusqu'à présent, nous avons considéré que la matrice de génotypes $\boldsymbol{G}$ était observée et permettait de construire directement la matrice $\boldsymbol{X}$. Dans le cas d'une approche par imputation, seuls les génotypes d'un *panel* d'imputation sont observés à tous les marqueurs, les individus constituant l'étude n'étant génotypés qu'à un sous-ensemble de marqueurs. Si l'on note la matrice de génotypes observée $\boldsymbol{G_{obs}}$, notre approche a consisté à intégrer dans notre inférence une étape d'échantillonage dans la distribution $P(\boldsymbol{G}|\boldsymbol{G_{obs}})$. Nous avons testé deux modèles pour effectuer cet échantillonage, soit en utilisant le modèle de Scheet et Stephens (2006) soit celui de Li et Stephens (2003). A partir de cette échantillonage, l'intégration dans l'analyse d'association consiste , dans le cas du prior D_1 à intégrer à l'algorithme MCMC une étape d'échantillonage dans la distribution *a posteriori* $P(\boldsymbol{G}|\boldsymbol{G_{obs}}, \boldsymbol{y})$, ou, dans le cas du prior D_2 à moyenner les calculs sur un échantillon issu de $P(\boldsymbol{G}|\boldsymbol{G_{obs}})$. Par la suite Guan et Stephens (2008) ont montré qu'utiliser les moyennes *a posteriori* des génotypes imputés était en général suffisant pour reproduire les résultats obtenus par ré-échantillonage tout en offrant des temps de calculs bien plus réduits.

Intérêt de la méthode

Puissance et robustesse vis-à-vis du déterminismes génétique Du fait de la modélisation des effets génétiques, autorisant la dominance, nous avons pu montrer que la puissance de détection était peu affectée par le déterminisme génétique sous-jacent. Contrairement à notre approche, des tests classiques doivent faire un choix en terme de modélisation : soit ils autorisent la dominance (test génotypique à 2 degrés de liberté) soit non (test allélique à un degré de liberté). Et ce choix a des conséquences sur la puissance de chacun des tests : le test allélique sera moins puissant dans le cas où la dominance est forte et inversement. Le Bayes Factor que nous avons proposé permet une plus grande souplesse dans la modélisation génétique et s'avérait le plus puissant dans la grande majorité des cas de figure. En particulier, notre approche se montrait largement plus puissante lorsque nous simulions plusieurs variants causaux dans une région génomique.

Un résultat plus surprenant est apparu dans les situations où tous les variants étaient génotypés : notre statistique (le BF) était plus puissante qu'un test basé sur les p-valeurs. Ceci mettait en évidence un défaut intrinsèque à l'utilisation des p-valeurs en tant que mesure de significativité d'une hypothèse. En effet, l'interprétation d'une p-valeur doit dépendre du niveau d'informativité du test effectué, spécifiquement de sa distribution attendue sous l'hypothèse alternative, qui est en général inconnue. Ceci n'est pas le cas des Bayes Factor dont l'interprétation ne dépend pas de la taille de l'étude.

Performance de l'approche par imputation Tout d'abord nous avons pu montrer que l'inférence obtenue après imputation permettait de retrouver les résultats obtenus si l'ensemble des données génétiques étaient connues : la puissance de détection et les erreurs d'estimation des effets étaient similaires. Nous avons pu en revanche constater que pour les variants rares, l'approche par imputation ne permettait pas de restaurer entièrement ce qui pouvait être obtenu lorsque tous les polymorphismes étaient connus. Ceci provient du fait que des variants rares sont souvent en déséquilibre de liaison incomplet avec les autres marqueurs. Cependant dans certaines situations simulées, un variant rare était effectivement retrouvé par l'approche par imputation alors qu'il ne l'était pas lorsque seuls les SNPs génotypés étaient inclus dans l'analyse. Il s'agit de cas où les génotypes du SNP causal pouvaient être imputés correctement à l'aide des haplotypes aux autres marqueurs. Bien que rares, j'ai rencontré une situation de ce type dans notre analyse sur le déterminisme du taux de recombinaison chez le mouton (cf. Section 2.3).

Conclusion

La méthode que j'ai développé lors de mon post-doctorat a par la suite été améliorée en particulier pour l'application à des analyses d'association incluant tout le génome (je n'ai pas participé à ces travaux). Pour permettre cette extension, deux principales adaptations ont été nécessaires. Tout d'abord Guan et Stephens (2011) ont étendu l'approche multi-SNP en ajoutant une procédure d'estimation des hyper-paramètres p_0 et σ_a^2, ce qui impliquait de définir de nouveaux *a priori* pour ces paramètres et un nouvel algorithme MCMC. Puis, il est apparu important de prendre en compte qu'une partie du déterminisme génétique du caractère pouvait être polygénique *i.e.* que l'ensemble des SNPs du génome ont des effets faibles sur le caractère, conduisant à une matrice

résiduelle de la même forme que dans l'équation (2.2). Cette extension, décrite par Zhou *et al.* (2013), est appelée Bayesian Sparse Linear Mixed Model (BSLMM). C'est la méthode que nous avons utilisé pour notre analyse du déterminisme génétique du taux de recombinaison chez le mouton.

2.3 Analyse du déterminisme du taux de recombinaison chez le mouton

Dans le cadre de notre étude sur la recombinaison ovine **(Petit *et al.*, 2017)**, nous avons conduit une étude d'association pour un phénotype de recombinaison : le nombre de crossing-overs par méiose. En effet, notre jeu de données comprenait 345 béliers pour lesquels nous disposions d'au moins deux descendants (méioses). Ainsi, nous disposions pour ces animaux de données répétées sur un nombre de crossing overs ayant eu lieu lors de leur méiose. Nous avons tout d'abord ajusté ce phénotype pour prendre en compte des effets fixes (mois de production de la semence) et âge de l'individu à la date d'insémination. Pour cela nous avons estimé le modèle suivant :

$$y_{so} = \mu + \boldsymbol{x}_{so}\boldsymbol{\beta} + a_s + u_s + e_{so} \tag{2.8}$$

où s indice le bélier, o le descendant, y_{so} est le nombre de crossing overs au cours de la méiose de s à o, $\boldsymbol{x}_{so}$ la ligne de la matrice d'incidence des effets fixes, $\boldsymbol{a} \sim \mathcal{N}(0, \boldsymbol{I_S}\sigma_a^2)$, $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{A}\sigma_s^2)$ et $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{I_N}\sigma_e^2)$. L'effet a_s est l'effet bélier individuel, u_s est la valeur génétique additive du bélier et $\boldsymbol{A}$ est la matrice d'apparentement entre individus estimée à partir de leur pedigree. De cette analyse nous avons obtenu une estimation de l'héritabilité du nombre de crossing-overs par méiose de 0.23 ± 0.02. Dans la suite de notre analyse, nous avons utilisé les prédictions des effets u_s comme phénotype pour les 345 béliers.

Nous avons utilisé les individus Lacaune génotypés sur la puce HD **(Rochus *et al.*, 2018)** comme panel d'imputation et imputés les 345 individus pour 534.000 marqueurs. Nous avons ensuite conduit une analyse d'association en utilisant la méthode de Zhou *et al.* (2013) en particulier. Cette analyse a mis en évidence la présence de deux QTLs majeurs et 3 régions suggestives (Figure 2.3). Ces 5 QTLs expliquaient dans leur ensemble près de 40% de la variance génétique additive, les 60% restant étant attribués à des effets polygéniques. Parmi les deux QTLs majeurs, l'un d'entre eux n'était révélé qu'avec les données imputées. L'autre QTL majeur co-localisait avec un gène connu pour affecter ce phénotype de recombinaison chez d'autres espèces, le gène RNF212. Nous avons alors reséquencé partiellement ce gène et pu détecter un SNP en très forte association avec la caractère se situant dans un de ses introns. Ces deux régions incluant des QTLs majeurs avaient été détectées par le même type d'étude dans la population Soay (Johnston *et al.*, 2016). Cependant, dans cette population le QTL situé près du gène RNF212 ne s'exprimait que chez les femelles. Ainsi, il n'est pas clair que ce soit le même QTL mais si ça l'est son effet a évolué de façon divergente dans ces deux populations. En ce qui concerne l'autre QTL majeur, l'identification d'un gène candidat n'a pas été possible : la région la plus significative est riche en séquences répétées et plusieurs gènes proches sont des candidats fonctionnels potentiellement pertinents (REC8, RNF212B, HEI10).

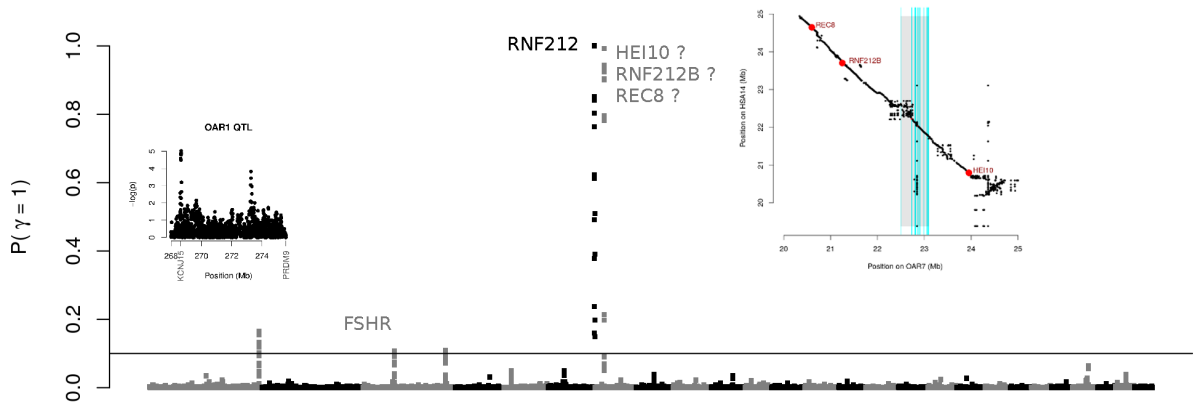


FIGURE 2.3 : Résultats de l'étude d'association pan-génomique pour le nombre de crossing-overs par méiose dans la population Lacaune mâle. L'axes des abscisses représente les marqueurs dans leur ordre sur le génome, l'axe des ordonnées est la probabilité a posteriori qu'un SNP soit un QTL. Les gènes candidats sont indiqués au dessus des régions candidate. Le panel de gauche représente une zoom du signal pour la régnion du chromosome 1. Le panel de droite présente l'alignement local du génome ovin (abscisses) et du génome humain (ordonnées) dans la région candidate du chromosome 7. Les SNPs candidats sont indiqués par les lignes bleues.

2.4 Conclusion

La détection de polymorphismes impliqués dans la variabilité des caractères complexes a connu des avancées spectaculaires issues d'une part des développements technologiques permettant d'interroger toujours plus en détail la variabilité génétique et d'autre part les méthodologies statistiques pour l'imputation de génotypes et la modélisation des effets génétiques. Au delà de leur intérêt premier pour l'étude d'un caractère particulier, la combinaison de multiples études offre aujourd'hui la possibilité de comprendre de manière plus globale le déterminisme génétique des caractères et leurs corrélations génétiques (e.g. Pickrell *et al.*, 2016). Dans l'espèce humaine, ce type d'analyse a mis en évidence d'une part que la plupart des caractères complexes ont une base extrêmement polygénique mais également qu'une grande proportion des SNP avaient des effets pléiotropes sur de nombreux caractères. Ces observations ont conduit le groupe de Jonathan Pritchard à suggérer un modèle de déterminisme "omnigénique" où l'ensemble des variants d'expression sont susceptibles de contribuer à un caractère (Boyle *et al.*, 2017 ; Liu *et al.*, 2019). Les espèces agronomiques, et en particulier les animaux d'élevage sont potentiellement un bon modèle pour contribuer à étudier cette question dans un contexte différent. Leur histoire évolutive a été marquée par la création d'une forte structuration génétique associée à des goulots d'étranglements et un processus de domestication et de sélection aboutissant à des individus exprimant des valeurs de caractères inconnues chez leurs ancêtres sauvages. Ces perturbations historiques ont laissé des traces sur le génome qui, si l'on est en mesure de les identifier et les interpréter, peuvent nous aider à comprendre le déterminisme génétique de cette évolution phénotypique.

3. Du phénotype au génotype : génétique de l'adaptation

Dans le chapitre précédent, nous avons vu comment le déterminisme génétique des caractères complexes pouvait être révélé à l'aide de modèles statistiques d'association génotype / phénotype. Ces modèles sont basés sur le lien fonctionnel entre l'information portée par l'ADN, son interprétation par la machinerie cellulaire pour aboutir à une expression d'un phénotype mesurable sur un individu. Lorsque l'on se place d'un point de vue évolutif, la sélection des individus sur des caractères influençant la capacité reproductive des individus (fitness) va provoquer une modification de la diversité génétique de la population au niveau des locus impliqués dans leur déterminisme. Une manière d'identifier ces locus est donc de rechercher dans le génome les régions qui semblent avoir évolué sous sélection : des *signatures de sélection*.

Cette approche est particulièrement adaptée à l'étude des populations d'animaux domestiques. En effet, l'histoire évolutive de ces populations est marquée par un processus continu de sélection depuis leur domestication, en passant par la colonisation de nouveaux milieux et jusqu'à l'établissement de races d'élevage. Conjointement à ce processus *adaptatif*, cette évolution procède également de phénomènes d'évolution *démographiques*. Les processus démographiques correspondent à des évènements de migration (colonisation de nouveaux milieux), de variation de taille (expansion ou rétraction) ou de croisement entre populations différenciées. Les nouveaux outils moléculaires (génotypage et séquençage haut-débit) offrent une information riche pour permettre la caractérisation de l'histoire démographique des populations et des facteurs génétiques (mutations) impliqués dans la réponse à la sélection.

Dans cette partie de mon rapport, je vais décrire les méthodes statistiques permettant d'exploiter ces données pour la détection de signatures de sélection, intégrant une modélisation des processus démographiques. Je présenterai les méthodes que j'ai contribué à développer ainsi que les résultats obtenus sur des jeux de données réelles. Ces travaux ont été effectué dans le cadre de l'encadrement d'une étudiante de master (Claire Oget), de trois thèses (Maria-Ines Fariello, Christina Rochus, Jason Lapeyronie), de l'accueil d'un doctorant étranger (Mahmood Gholami) et du travail d'un chercheur en post-doctorat encadré par des membres de mon équipe (Maxime Bonhomme).

3.1 Principe de détection des signatures de sélection

Pour comprendre le principe de la recherche de mutations sélectionnées il faut tout d'abord comprendre comment la sélection sur un caractère (phénotype) entraîne une modification de la diversité génétique. La figure 3.1 illustre schématiquement la réponse à la sélection au niveau d'une mutation en une seule génération.

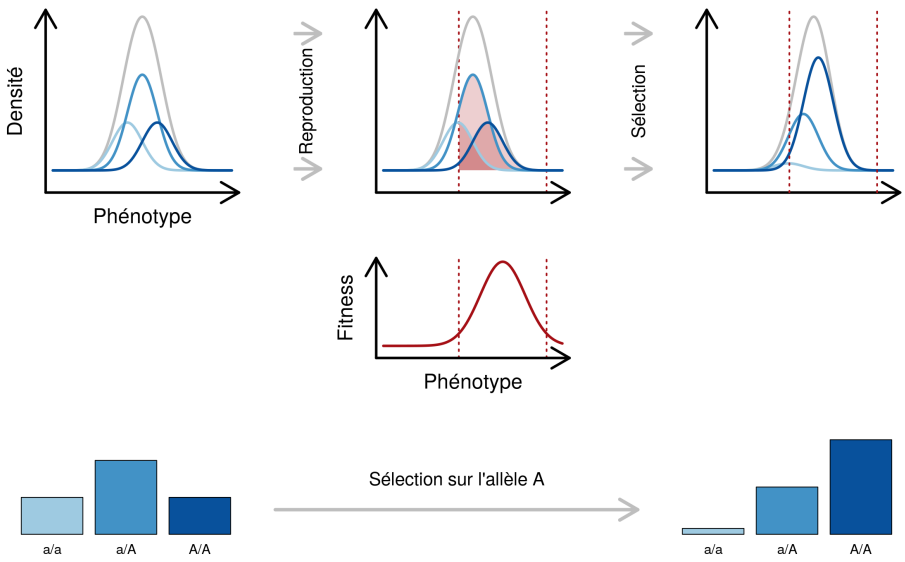


Figure 3.1 : Illustration de l'effet de la sélection sur une mutation génétique participant à l'élaboration du phénotype sélectionné.

On considère une population dans laquelle un caractère est en partie déterminé par un variant génétique. Ce variant admet deux allèles a et A et la valeur du caractère d'un individu la population dépend en partie de son génotype : les individus porteurs d'un ou deux allèles A ont des valeurs plus élevées pour le caractère. Lors de la reproduction des individus d'une génération à l'autre, la capacité reproductive des individus (fitness) dépend de leur valeur pour le phénotype. Dans le cas de la figure 3.1 les individus porteurs d'une valeur élevée pour le caractère ont une fitness plus élevée. Comme ces individus sont aussi plus souvent porteurs de l'allèle A, cet allèle va avoir plus de chance d'être transmis à la génération suivante et la fréquence de l'allèle A dans la population va augmenter.

Cet effet illustré sur une génération va se cumuler au cours du temps et dans une population sélectionnée, les allèles favorables pour les caractères de fitness vont voir leur fréquence augmenter graduellement, potentiellement jusqu'à envahir la population. Dans une population qui ne subit pas la même sélection, ce même allèle évoluera uniquement du fait des processus démographiques selon une trajectoire aléatoire. La figure 3.2 illustre la trajectoire temporelle d'une petite région génomique dans deux populations de même origine mais subissant ou non une pression de sélection. Dans la population sélectionnée, l'allèle affectant les caractères de fitness (en rouge) voit sa fréquence augmenter au cours du temps pour finalement fixer dans la population alors qu'il finira par disparaître dans l'autre population. Cette figure permet d'illustrer les trois grands types d'approche utilisées pour détecter les régions génomiques ayant été affectées par la sélection.

Le premier type d'approches se base sur la modification de la diversité locale autour d'une mutation sous sélection (A. en bleu sur la figure 3.2). Du fait de la liaison entre locus portés par le même chromosome, la sélection d'un allèle favorable va entraîner l'augmentation en fréquence des allèles qui lui sont associé initialement (en noir sur la figure) et conduire à homogénéiser les ressemblances entre chromosomes dans la population. Certaines méthodes de détection vont donc rechercher à identifier des régions génomiques présentant ces caractéristiques. Elles vont

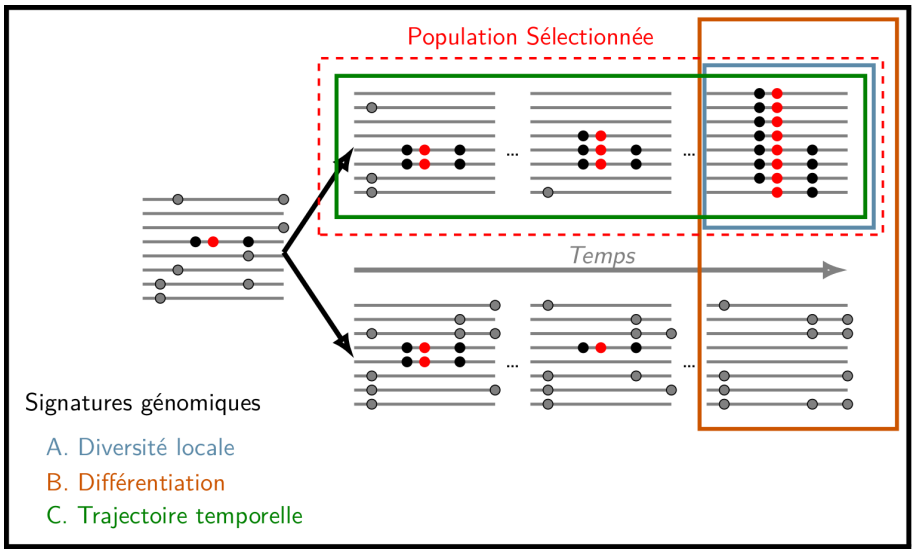


FIGURE 3.2 : Trajectoire évolutive d'un allèle dans une population sélectionnée ou neutre et grands types de signatures de sélection.

être surtout efficaces quand la mutation a été complètement fixée dans la population, c'est à dire qu'elles identifient plutôt des évènements de sélection anciens.

Le deuxième type d'approches va exploiter la disparité des évolutions entre populations ayant subi des pressions de sélection différentes (B. en orange sur la figure 3.2). Dans les régions ayant répondu à la sélection dans une population mais pas dans une autre, les différences des patrons de diversité locales entre populations vont être accentuées et pourront être détectées par contraste avec les autres régions génomiques. L'intérêt de ce type d'approches est qu'elles vont permettre de détecter les effets de la sélection même si les allèles favorables n'ont pas fixé. Elles vont donc être en mesure de détecter des effets d'une sélection récente. En revanche, si les populations sont séparées depuis longtemps, la différentiation globale entre population peut être telle qu'il sera difficile de séparer les signatures de sélection du bruit de fond global.

Le dernier type d'approches va exploiter les situations dans lesquelles des données génomiques temporelles sont disponibles. Ces approches vont alors rechercher à détecter les mutations dont la fréquence allélique augmente de manière non aléatoire au cours du temps (C. en vert sur la figure 3.2). Bien que ce type de données soient encore aujourd'hui assez rares, il est appelé à devenir plus fréquent dans l'avenir en particulier dans trois contextes très différents. Tout d'abord, certaines populations sont aujourd'hui génotypées en routine pour des objectifs de sélection génomique. Le temps passant, les données génomiques s'accumulent et les trajectoires de réponse à la sélection vont devenir disponibles. Ensuite, l'essor des technologies permettant le séquençage d'ADN à partir de fossiles va permettre d'avoir accès à des données génétiques remontant jusqu'à plusieurs milliers d'années dans le passé, c'est à dire jusqu'aux origines de la domestication. Enfin, de nombreux dispositifs de sélection expérimentale ont été mis en place depuis plusieurs années, en particulier au département de génétique animale de l'INRA. Ces dispositifs consistent souvent à sélectionner de manière divergente des populations issues d'un même groupe de fondateurs. Ce type de dispositif peut permettre d'identifier les mutations contribuant à la réponse à la sélection en détectant leurs trajectoires temporelles extrêmes.

Dans le cadre de mon travail, j'ai contribué principalement au développement de méthodes basées sur les signatures de différentiation entre populations. Ce type d'approches est particulièrement bien adapté aux espèces d'animaux d'élevage qui sont souvent structurées en populations isolées génétiquement (*e.g.* en races gérées par les éleveurs). Plus récemment j'ai participé à l'évaluation de méthodes basées sur l'exploitation de données génétiques temporelles (cf. Chapitre 4).

3.2 La méthode FLK

Lors de mon arrivée au laboratoire j'ai contribué au développement d'une méthode de détection de sélection proposée par Magali San Cristobal et Claude Chevalet dans le cadre du travail post-doctoral de Maxime Bonhomme **(Bonhomme *et al.*, 2010)**. Le principe de cette méthode est d'établir un modèle neutre (sans sélection) pour la distribution conjointe des fréquences alléliques à un locus dans plusieurs populations. Ce modèle est ensuite ajusté sur les données et une statistique d'ajustement [1] est calculée à chaque marqueur. Elle peut ensuite être utilisée pour tester le modèle neutre. En cas de rejet du modèle neutre, le locus testé est déclaré sous sélection. Je vais décrire dans cette section le principe de cette méthode et son avantage par rapport aux approches encore couramment utilisées pour la détection de signatures de sélection basées sur la différenciation entre populations. Cette présentation est sensiblement différente de celle que l'on peut trouver dans la publication originelle.

Modèle neutre d'évolution dans une population

Avant d'introduire le modèle multivarié pour un vecteur de fréquences alléliques observées aujourd'hui, il est intéressant de considérer un modèle approché de l'évolution de la fréquence allélique au cours du temps dans une population. Nous voulons modéliser la loi de la fréquence allélique à un locus dans une population de taille N diploïdes évoluant sous le modèle de Wright-Fisher sans sélection ni mutation. Sous ces hypohèses, si l'on note p_t la fréquence allélique à la génération t on montre que :

$$\begin{aligned} \mathbb{E}(p_t) &= p_0 \\ \mathbb{V}(p_t) &= \Big(1 - (1 - \frac{1}{2N})^t\Big) p_0(1 - p_0) \end{aligned} \tag{3.1}$$

où p_0 est la fréquence au début du processus. Ces équations nous disent que l'espèrance de la fréquence allélique est constante sous neutralité et que la variance augmente au cours du temps. Il faut entendre cette variance comme la variance du processus évolutif qui partirait d'une fréquence p_0. Ici il est intéressant de s'attarder sur le terme de de la variance qui ne dépend pas de p_0 : $\Big(1 - (1 - \frac{1}{2N})^t\Big)$, noté F_t **(Bonhomme *et al.*, 2010)**, et de son interprétation génétique.

On peut tout d'abord noter que pour $t \to \infty$ ce terme tend vers 1 et la variance asymptotique est donc $p_0(1 - p_0)$ ce qui correspond bien à la situation où l'allèle a fixé en 1 (probabilité p_0)

1. *goodness-of-fit* en anglais

ou 0 (probabilité $(1 - p_0)$) (la fréquence p_∞ suit alors une loi de Bernoulli de paramètre p_0). Ensuite on peut remarquer que ce terme est aussi la probabilité que deux allèles tirés au hasard dans la population aient un ancêtre commun plus récent que la génération 0. Ainsi cette équation met en lumière la relation entre la distribution des fréquences alléliques dans une population et le processus évolutif qui l'a généré. Quand $t \ll 2N$, $F_t \approx t/2N$, ce qui montre que la variance augmente de façon approximativement proportionelle aux nombres de générations et inversement proportionelle à la taille de population. Ainsi, F_t mesure la quantité de dérive accumulée depuis la génération initiale.

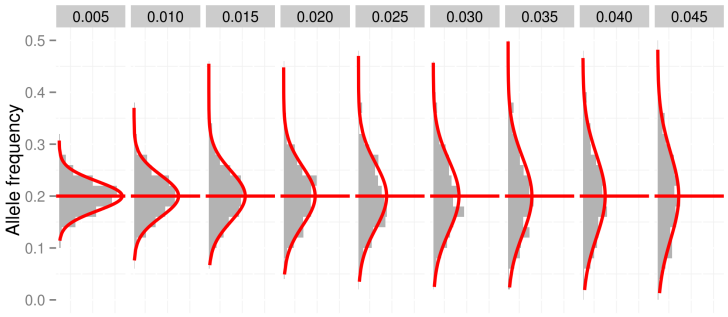


FIGURE 3.3 : Approximation du processus de Wright-Fisher sous dérive pure avec une loi normale. En gris : distribution simulée sous Wright-Fisher. En rouge : approximation par une loi normale de paramètres égaux aux moments du processus de Wright-Fisher.

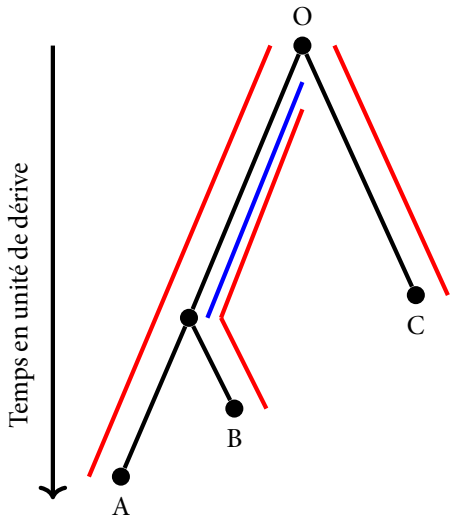
Le jeu d'équations (3.1) établit les deux premiers moments du processus de Wright-Fisher sous dérive pure. Ces moments peuvent ensuite être utilisés pour établir un modèle de la distribution des fréquences alléliques. Une possibilité initialement suggérée par Nicholson *et al.* (2002) est d'utiliser la loi normale comme approximation de la distribution de p_t. Cette approximation peut paraître inadaptée dans la mesure où elle a pour support $\mathbb{R}$ et non $[0, 1]$ mais dans les conditions $t \ll 2N$ et pour $p_0 \in [\epsilon, 1 - \epsilon]$ avec ϵ pas trop petit, il s'agit d'une bonne approximation (Figure 3.3). Par ailleurs, l'intérêt de l'utilisation de cette famille de loi pour la modélisation est sa capacité à être étendue au cas multivarié que nous allons maintenant présenter.

Modèle d'évolution dans plusieurs populations

Dans le cadre de la méthode FLK, le modèle neutre précédent est étendu à plusieurs populations en considérant d'une part que ces populations sont toutes issues d'une même population ancestrale et d'autre part que la divergence depuis cette population ancestrale s'est déroulée par dérive pure et séparation. Il est fait l'hypothèse qu'il n'y a pas de migration entre populations une fois qu'elles sont séparées. L'histoire évolutive des populations est alors représentée par un arbre dont la racine est la population ancestrale (Figure 3.4). En suivant le même raisonnement que pour le modèle d'une seule population, on peut décrire la loi conjointe des fréquences alléliques dans plusieurs populations comme une loi normale multivariée dont la matrice de covariance dépend des longueurs de branches dans l'arbre. Le modèle multipopulationnel est alors :

$$\mathbf{p} \sim \mathcal{N}(p_0\mathbf{1}, \mathbf{F}p_0(1 - p_0)) \tag{3.2}$$

où $\boldsymbol{F}$ est une matrice décrivant les covariances entre fréquences alléliques (Figure 3.4) : dans cette illustration, les populations A et B se sont séparées dans un deuxième temps et leurs fréquences alléliques p_A et p_B ont une covariance proportionelle à la quantité de dérive accumulée avant leur séparation.



Modèle Gaussien multivarié

$\mathbf{p} \sim \mathcal{N}(p_0\mathbf{1}, \mathrm{F}p_0(1-p_0))$

$Var(p_X) = f_X\, p_0(1-p_0)$

$Cov(p_A, p_B) = f_{AB}\, p_0(1-p_0)$

$$\mathbf{V}(\mathbf{p}) = \begin{bmatrix} f_A & f_{AB} & 0 \\ f_{AB} & f_B & 0 \\ 0 & 0 & f_C \end{bmatrix} p_0(1-p_0)$$

$$\mathbf{V}(\mathbf{p}) = \mathrm{F}\, p_0(1-p_0)$$

Figure 3.4 : Exemple d'un arbre phylogénétique de trois populations (A, B et C). La racine de l'arbre correspond à la population ancestrale (O) et le temps se déroule de haut en bas. Les longueurs de branche de l'arbre sont mesurées en unité de dérive qui dépend de la taille efficace de chaque population. Les branches internes correspondent à des populations ancêtres des populations actuelles.

L'intérêt de cette approche est qu'elle permet de prendre en compte les évolutions démographiques passées pour modéliser les observations contemporaines. Spécifiquement les populations peuvent avoir des tailles efficaces différentes et des relations de parentés spécifiques. Sa limite principale réside dans le fait qu'elle ne prend pas en compte la possibilité de migration entre populations. Ceci peut cependant être pris en compte dans une certaine mesure par une adaptation de la procédure d'estimation de la matrice $\boldsymbol{F}$ que nous allons désormais décrire.

Estimation du modèle FLK et test de neutralité

Le modèle (3.2) contient deux paramètres à estimer : la fréquence dans la population ancestrale p_0 (spécifique à chaque marqueur) et la matrice de covariance $\boldsymbol{F}$. Si $\boldsymbol{F}$ est connue, un estimateur de p_0 est l'estimateur des moindres carrés généralisés :

$$\hat{p_0} = \frac{\mathbf{1}^t \boldsymbol{F}^{-1} \boldsymbol{p}}{\mathbf{1}^t \boldsymbol{F}^{-1} \mathbf{1}} \tag{3.3}$$

L'estimation du modèle (3.2) requiert donc en fait uniquement de savoir estimer la matrice de covariance $\boldsymbol{F}$. La matrice $\boldsymbol{F}$ traduit la démographie des populations et décrit donc les covariances

de tous les locus du génome. Son estimation s'effectue en combinant l'information issue de plusieurs locus, idéalement des locus neutres. En pratique elle est estimée en considérant un grand nombre de locus du génome sous l'hypothèse implicite que la grande majorité d'entre eux ne sont pas sous sélection. La méthode d'estimation de $\boldsymbol{F}$ de **Bonhomme *et al.* (2010)** consiste à (i) estimer les distances entre paires de populations *mesurées en unité de dérive*, (ii) reconstruire l'arbre des populations à partir de cette matrice en utilisant l'algorithme de Neighbour-Joining (Saitou et Nei, 1987) et (iii) enraciner l'arbre des populations pour identifier la population ancestrale.

Pour estimer les distances entre paires de populations, **Bonhomme *et al.* (2010)** utilisent la distance de Reynolds (Reynolds *et al.*, 1983) que Laval *et al.* (2002) avaient montré comme la plus adaptée pour des populations évoluant sous dérive pure. Pour l'enracinement, ils suggèrent d'utiliser un groupe externe fortement différencié. Dans la version actuelle de la méthode, il est aussi possible d'enraciner l'arbre sans groupe externe en recherchant la position de la racine qui minimise la différence entre l'hétérozygotie observée dans chacune des populations et celle prédite par les longueurs de branches de l'arbre (Claude Chevalet, pers. comm.).

Pour prendre en compte la possibilité d'admixture entre populations, il est possible d'ajouter une étape supplémentaire dans la procédure d'estimation de $\boldsymbol{F}$: à partir de l'estimateur initial de $\boldsymbol{F}$, on estime à chaque marqueur le paramètre p_0 correspondant (Equation (3.3)) puis on construit les résidus $\boldsymbol{r} = (\boldsymbol{p} - \hat{p_0})/\sqrt{\hat{p_0}(1-\hat{p_0})}$. La nouvelle matrice $\boldsymbol{F}$ est alors la matrice de covariance empirique $\boldsymbol{r}\boldsymbol{r}^{\boldsymbol{t}} + \boldsymbol{\epsilon}$ avec $\boldsymbol{\epsilon}$ une matrice diagonale ayant des termes très petits pour stabiliser l'inverse $\boldsymbol{F}^{-1}$. Cette procédure consiste donc à obtenir des covariances entre populations qui sont construites sous une contrainte d'évolution en arbre tout en autorisant des déviations à ce modèle.

Une fois que la matrice $\boldsymbol{F}$ est estimée, il est possible de construire une statistique d'ajustement du modèle à chaque marqueur. Cette statistique nommée T_{F-LK} par **Bonhomme *et al.* (2010)** s'écrit :

$$T_{F-LK} = B\,(\boldsymbol{p} - \hat{p_0}\mathbf{1})^t \frac{\boldsymbol{F}^{-1}}{\hat{p_0}(1-\hat{p_0})}(\boldsymbol{p} - \hat{p_0}\mathbf{1}) \tag{3.4}$$

où B est un terme de correction qui provient du fait que $\mathbb{E}(\hat{p_0}(1-\hat{p_0})) \neq p_0(1-p_0)$. Dans le cadre de modèles linéaires, cette statistique est la déviance du modèle. Sous l'hypothèse où les fréquences alléliques peuvent être modélisées par une loi normale, cette statistique suit une distribution du χ^2 à $n-1$ degrés de liberté où n est le nombre de populations considéré. Ainsi le modèle neutre peut être testé à chaque marqueur du génome par un test paramétrique rapide à calculer. De ce fait, la méthode FLK fournit un moyen efficace de tester le modèle neutre sur des jeux de données de grande dimension. Par ailleurs, à travers la modélisation en arbre ce test permet de corriger les défauts du test dont il s'inspire et proposé par Lewontin et Krakauer (1973).

3.3 Méthode haplotypique pour la détection de signatures de sélection : hapFLK

Dans le cadre de la thèse de Maria-Ines Fariello nous avons mis au point une extension de la méthode FLK qui permet de combiner l'information sur plusieurs marqueurs adjacents pour dé-

tecter des signatures de sélection par différentiation : la méthode hapFLK (Fariello *et al.*, 2013). Le principe est d'exploiter un modèle multipoint des associations alléliques (déséquilibre de liaison, abbr. DL) pour transformer les génotypes bialléliques aux SNPs en génotypes multialléliques. Ces génotypes sont ensuite combinés dans une statistique d'ajustement de modèle neutre basée sur les mêmes principes que FLK. Dans cette section, je présenterai d'abord le modèle multipoint utilisé puis la méthode hapFLK et ses performances évaluées sur simulation et données réelles.

Modèle multipoint du déséquilibre de liaison

Pour modéliser les associations alléliques non aléatoires entre locus proches, nous avons utilisé le modèle développé par Scheet et Stephens (2006) (abbrv. SSM). Je vais ici décrire le principe de la méthode sans entrer dans les détails qui peuvent être trouvés dans la référence précédente. Le modèle SSM est un modèle de markov caché (Hidden Markov Model, abbr. HMM) dont les états cachés sont des haplotypes "templates" aussi appelés "clusters". Les observations sont soit des haplotypes résolus soit des génotypes (paires d'haplotypes non résolus). Pour décrire le principe du modèle il est plus simple de se placer dans le premier cas, ce que je ferai ici, mais il est important de noter que l'utilisation du modèle SSM peut se faire sur des données non phasées ce qui est un de ses avantages pour l'utilisation que nous en avons faite. Pour caractériser un HMM, il faut décrire d'une part les probabilité de transition entre états cachés et d'autre part les probabilités d'émission, *i.e.* les probabilités des observations conditionnellement à l'état caché.

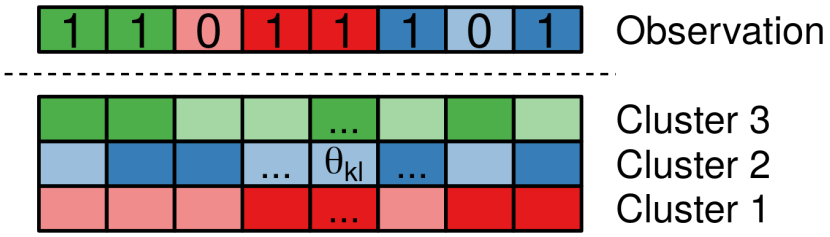


Figure 3.5 : Illustration de la modélisation d'un haplotype observé sous le modèle SSM. Les états cachés sont des vecteurs de probabilité d'émission (θ_{kl}). Un haploptype observé est modélisé comme une mosaique d'états cachés. Ici une mosaique possible est représentée.

La Figure 3.5 représente schématiquement le principe de modélisation d'un haplotype du modèle SSM. Si l'on appelle h_l l'allèle observé au locus l pour l'haplotype considéré et z_l de cluster dont provient h_l au locus l, la probabilité d'émission du HMM est $P(h_l = 1|z_l = k) = \theta_{kl}$: les états cachés sont des vecteurs de probabilité d'émission θ_{kl} . Les probabilités de transition $P(z_{l+1} = k'|z_l = k)$ pour $k \neq k'$ pénalisent le changement de cluster entre deux marqueurs. Le choix du nombre de clusters doit être fait en amont. Dans l'utilisation du modèle SSM pour le phasage une procédure de cross validation est utilisée pour le déterminer.

L'estimation de ce modèle se fait par maximum de vraisemblance à l'aide de l'algorithme de Baum-Welsh en utilisant un ensemble d'observations (haplotypes ou génotypes). Ceci conduit typiquement à des estimations $\widehat{\theta_{kl}}$ prochent de 0 ou 1 ce qui correspond à la situation où les états cachés sont des haplotypes "templates". Les probabilités de transition peuvent aussi être estimées,

ou alors fixées à des valeurs faibles pour pénaliser les changements trop fréquents de cluster. Dans le cadre de notre utilisation de ce modèle, nous estimons le modèle sur l'ensemble des individus (toutes populations confondues) puis nous calculons à chaque marqueur les probabilités *a posteriori* $P(z_l = k|\boldsymbol{h})$. Ensuite, pour chaque population nous calculons la "fréquence" du cluster k (p_{kl}) comme la moyenne des probabilités *a posteriori* sur l'ensemble des individus de la population. Scheet et Stephens (2006) avaient remarqué que la procédure d'estimation par maximum de vraisemblance de ce modèle conduisait systématiquement à un maximum local. Ils ont montré que pour faire de l'inférence avec ce modèle il était alors plus efficace de la moyenner sur plusieurs estimations (*i.e.* plusieurs maximums locaux) plutôt que d'utiliser l'estimation qui donne la meilleure vraisemblance. C'est ce que nous avons fait également pour la statistique hapFLK que je vais maintenant décrire.

Intégration de la modélisation du DL dans l'approche FLK : hapFLK

Après avoir estimé le modèle de Scheet et Stephens (2006) nous avons construit une statistique basée sur les mêmes principes que FLK, mais utilisant les fréquences de clusters par population **(Fariello *et al.*, 2013)**. Le principe est de calculer, pour chaque cluster k une déviance, similaire à (3.4) :

$$D_0^k = \frac{(\mathbf{p^k} - \mathbf{1}'\hat{p}_0^k)'F^{-1}(\mathbf{p^k} - \ 1'\hat{p}_0^k)}{\sigma^2} \tag{3.5}$$

La différence importante étant que le paramètre de variance σ^2 (i) n'est pas spécifié en terme de fréquence "ancestrale" p_0 et (ii) est le même pour tous les clusters. La statistique hapFLK est ensuite calculée comme :

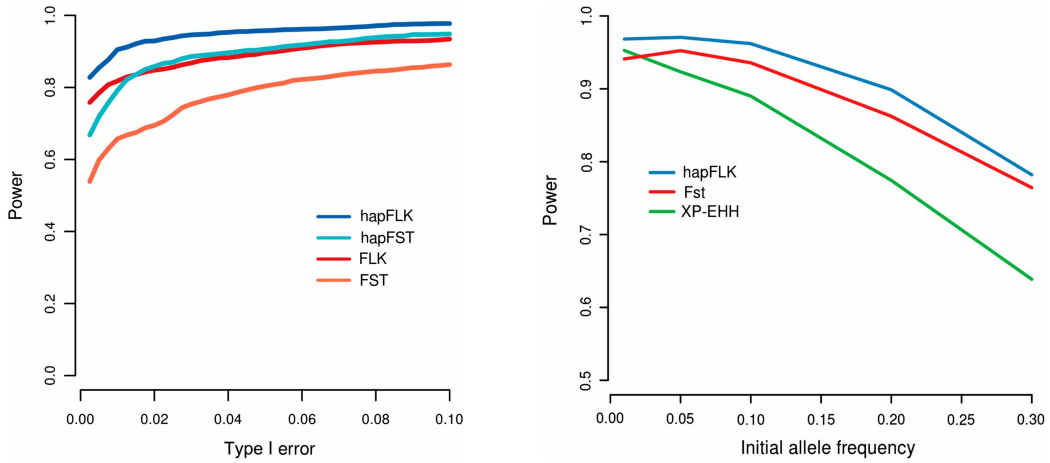
$$hapFLK = \sigma^2 \sum_k D_0^k \tag{3.6}$$

La composante de variance n'est donc pas estimée et la statistique hapFLK est de plus moyennée sur plusieurs estimations du modèle SSM (voir ci-dessus). Ainsi sa distribution asymptotique n'est pas connue mais peut être estimée à partir des données. Dans notre analyse des données du projet SheepHapMap **Fariello *et al.* (2013, 2014)**, la statistique présentait une distribution proche d'une loi normale. Nous avons donc utilisé une méthode d'estimation robuste aux outliers pour en estimer les paramètres. Plus tard, nous avons pris en compte le fait que la statistique, en tant que déviance, devait suivre une loi proportionelle à un χ^2 à (n-1)(K-1) degrés de liberté. Pour déterminer la constante σ^2 nous avons proposé d'effectuer une régression des quantiles observés sur les quantiles théoriques **(Boitard *et al.*, 2016)**.

Performance du test hapFLK

Pour déterminer les performances du test hapFLK par rapport aux approches classiques, nous avons effectué une série de simulations variant dans le nombre de populations considérées, la présence ou non d'évènements non pris en compte par le modèle démographique (admixture, goulots d'étranglement) et la fréquence initiale de l'allèle favorable. Nos résultats (*e.g.* Figure 3.6)

ont montré que le test hapFLK permettait souvent de gagner en puissance par rapport à des approches classiques (F_{ST}) ou d'autres approches utilisant l'information haplotypique (XP-EHH).



(a) Comparaison de différents tests (b) Puissance en fonction de la fréquence initiale

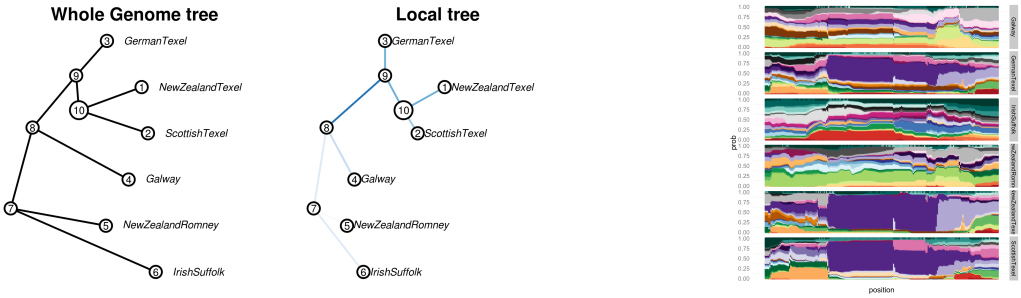
Figure 3.6 : Puissance du test hapFLK dans différents scénarios simulés.

Interprétation des signatures de sélection

Les tests FLK et hapFLK mesurent l'adéquation du modèle neutre a la différentiation locale des populations. En tant que tests d'ajustement ils ne nécessitent pas de spécifier un modèle alternatif d'adaptation. Cependant, dans la recherche de signatures d'adaptation il est souvent intéressant de pouvoir interpréter les signatures par exemple en cherchant à connaître les populations sélectionnées ou quels sont les allèles / haplotypes favorables. Dans le cadre du développement de la méthode hapFLK et lors de son application sur des jeux de données réelles, nous avons cherché à établir des méthodes d'annotation permettant de faciliter l'interprétation des signaux détectés. Pour essayer d'identifier les populations sélectionnées, nous avons proposé d'une part de reconstruire des arbres phylogénétiques locaux, en imposant la structure neutre et d'autre part une représentation graphique de la diversité génétique locale. Ces deux représentations sont illustrées sur la Figure 3.7.

3.4 Réponses des génomes des espèces animales domestiques à la sélection

Depuis la publication de la méthode hapFLK, j'ai participé à plusieurs analyses de jeux de données réelles chez différentes espèces animales, listées dans Tableau 3.1. Ces analyses varient (i) en terme de nombres de marqueurs disponibles : de puces de moyenne densité **(e.g. Fariello *et al.*, 2014)**, **(Bertolini *et al.*, 2018)** jusqu'à des données de reséquençage de génomes entiers **(e.g. Bouwman *et al.*, 2018; Alberto *et al.*, 2018)** et (ii) en terme de nombre de populations, depuis des jeux de



(a) Arbres phylogénétiques global (gauche) et local (droite). (b) Diversité haplotypique locale.

FIGURE 3.7 : Annotation de la signature de sélection autour du gène GDF8, codant pour la myostatine, dans des populations ovines. Ces annotations identifient clairement les populations Texel comme ayant été sélectionnées pour le même haplotype dans cette région.

Projet	Espèce(s)	# SNP	# pop.	Origine	Référence
SheepHapMap	Mouton	50 K	74	Monde	Fariello *et al.* (2014)†
Synbreed	Poulet	1 M	7	International	Gholami *et al.* (2015)†
1K Bull Genomes (v2)	Vache	15 M	4	International	Boitard *et al.* (2016)
OPA	Mouton	600 K	23	France	Rochus *et al.* (2018)†
1K Bull Genomes (v4)	Vache	20 M	18	Monde	Bouwman *et al.* (2018)
Nextgen	Chèvre + Mouton	33 M + 23 M	3 + 3	Iran & Maroc	Alberto *et al.* (2018)
AdaptMap	Chèvre	50 K	93	Monde	Bertolini *et al.* (2018)
AdaptMap (FR)	Chèvre	50 K	6	France	Oget *et al.* (2019)†

TABLE 3.1 : Analyses de données FLK / hapFLK. † Publications d'étudiants encadrés.

données ciblés sur un petit nombre de populations **(e.g. Gholami *et al.*, 2015 ; Boitard *et al.*, 2016 ; Oget *et al.*, 2019)** jusqu'à des jeux de données incluant une grande part de la diversité mondiale **(e.g. Fariello *et al.*, 2014 ; Bouwman *et al.*, 2018 ; Bertolini *et al.*, 2018)**. La plupart de ces jeux de données ont été créé par des consortiums internationaux lors du développement de la sélection génomique dans les espèces concernées. Il s'agissait alors de valider un nouvel outil de génotypage **(Fariello *et al.*, 2014 ; Bertolini *et al.*, 2018 ; Rochus *et al.*, 2018)** et/ou d'établir des panels représentatifs pour permettre l'imputation de génotypes dans des populations déjà génotypées **(Boitard *et al.*, 2016 ; Bouwman *et al.*, 2018 ; Rochus *et al.*, 2018)**. On peut noter le cas particulier du projet Nextgen qui avait des objectifs essentiellement académiques : étudier la domestication **(Alberto *et al.*, 2018)** et l'adaptation des populations locales chez les petits ruminants.

La démarche générale d'une analyse de données de ce type passe premièrement par l'établissement d'un modèle démographique neutre permettant d'expliquer le comportement de la plupart des polymorphismes du génome. Dans le cadre de l'utilisation de la méthode hapFLK, basée sur le principe de populations différenciées, il s'agit alors d'identifier les groupes d'individus homogènes génétiquement (les populations) et de s'assurer que les évènements d'admixture entre populations ne soient pas trop récents. Il convient aussi de s'assurer que les populations identifiées ne présentent pas des phénomènes de dérive forte comme ceux provoqués par des goulots d'étranglement. Une fois le modèle neutre établit, les tests hapFLK et FLK peuvent être appliqués. L'adéquation générale du modèle neutre peut alors être vérifiée en s'assurant que la statistique FLK suit, à l'exception de quelques outliers, sa distribution théorique. Pour la statistique hapFLK, cet ajustement est obtenu en recalibrant ses valeurs à l'aide d'estimation robuste des moments d'une loi normale (e.g. Fariello *et al.*, 2014) ou d'un χ^2 (e.g. Boitard *et al.*, 2016). Ceci étant fait, la détection de SNP ou d'haplotypes significatifs est obtenue en utilisant des approches classiques de correction pour les test multiples (e.g. Storey et Tibshirani, 2003). C'est un des intérêts de ces méthodes que de pouvoir avoir recourt à des approches éprouvées pour contrôler la significativité des résultats. Une fois les régions significatives identifiées, se pose la question de leur interprétation adaptative.

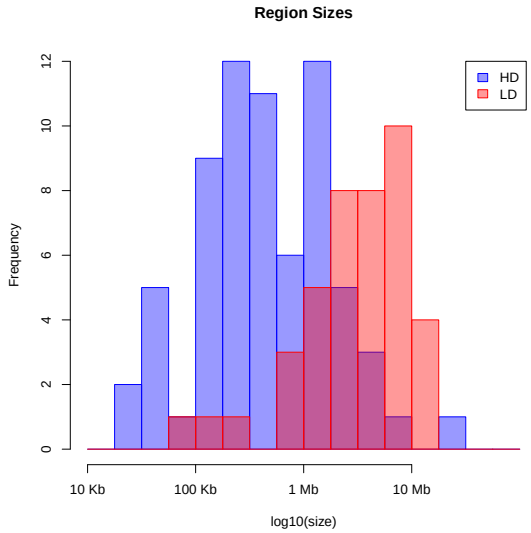


Figure 3.8 : Distribution des tailles (en échelle log) de signatures de sélection détectées avec une puce moyenne densité (LD, 60K) et avec une puce haute densité (HD, 600K)

Une fois les signatures significatives établies, la première interrogation qui survient est d'identifier le(s) gène(s) impliqué(s) voire les mutations adaptatives elle mêmes. De ce point de vue la densité de marqueurs utilisée est primordiale. Dans le cas d'études faisant appel à des puces de moyenne densité les régions génomiques identifiées sont parfois très grandes et le nombre de gènes candidats positionnels importants. Par exemple, dans le cas de l'analyse de Fariello *et al.* (2014) utilisant une puce de 50K marqueurs certaines régions contenaient plus de 200 gènes. Sur ce même type de populations, notre étude ultérieure (Rochus *et al.*, 2018) sur une puce de 600K marqueurs

réduisait d'un facteur 10 la taille des régions détectées (Figure 3.8). Nos analyses sur des données de séquences complètes **(Boitard *et al.*, 2016; Alberto *et al.*, 2018; Bouwman *et al.*, 2018)** ont confirmé cet augmentation de la précision de localisation avec la densité de marqueurs. Dans **Boitard *et al.* (2016)** sur données de séquence complète nous avons ainsi pu mettre en évidence que nous parvenions à identifier des mutations individuelles connues pour être impliquées dans le déterminisme génétique de caractères d'intérêt : MC1R pour la couleur de robe et PLAG1 pour la stature (Karim *et al.*, 2011 ; Bouwman *et al.*, 2018). Ceci nous laisse penser que d'autres mutations causales se trouvent parmi les mutations candidates que nous avons identifiées. Cependant un des résultats marquants de cette étude était que toutes les mutations candidates identifiées sauf une se trouvaient en dehors des régions codantes du génome. La validation de leurs effets adaptatifs sera donc d'autant plus compliquée. Une approche pour ce faire serait de combiner les résultats de recherche de locus adaptatifs avec ceux d'études d'association pan-génomique. Chez les bovins nous avons ainsi pu montrer que près de 50% des 163 mutations impliquées dans le déterminisme de la taille avaient aussi été impliqués dans la réponse à la sélection **(Bouwman *et al.*, 2018)**. Ceci est illustré sur la figure 3.9 qui montre la diversité haplotypique des races considérées dans deux signatures de sélection qui sont aussi des QTLs de stature.

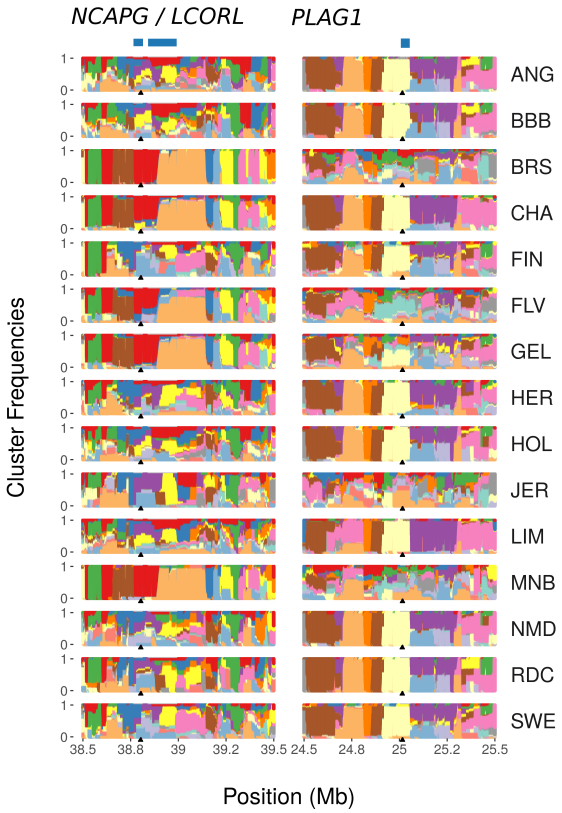


FIGURE 3.9 : Diversité haplotypique de 15 races bovines autour de 2 QTLs impliqués dans le déterminisme génétique de la taille. Les rectangles bleus indiques les gènes candidates et les triangles noirs les positions des SNP GWAS. Tirée de Bouwman *et al.* (2018)

Sur la figure 3.9, on peut voir que pour les deux régions, les haplotypes sélectionnés sont identiques pour toutes les populations ayant subi la sélection. Ce type de situation correspond soit à un allèle ancestral sélectionné de manière indépendante dans chaque population soit à un allèle fixé ancestralement. Dans les différentes études auxquelles j'ai participé cette situation n'est qu'un des cas rencontrés. Nous avons pu mettre en évidence des régions génomiques ou l'haplotype sélectionné différait entre populations, et donc qu'il s'agissait probablement d'allèles adaptatifs différents. Cette hétérogénéité allélique aux locus adaptatifs semble d'ailleurs assez commune [2]. Dans **Rochus *et al.* (2018)** nous avons d'ailleurs utilisé des méthodes statistiques spécifiques (Hormozdiari *et al.*, 2014) pour évaluer cette hétérogénéité allélique de manière quantitative et avons trouvé de nombreux cas compatibles avec cette hypothèse [3]. Pour l'une de ces signatures nous avons même validé cette hétérogénéité par reséquençage du gène candidat (MC1R [4]). Cependant, il faut noter que les situations où plusieurs haplotypes différents ont été sélectionnés dans des populations correspondent aussi aux cas pour lesquels les tests de type FLK et hapFLK sont très puissants. On peut donc s'attendre à les voir sur-représentés dans les régions déclarées comme significatives.

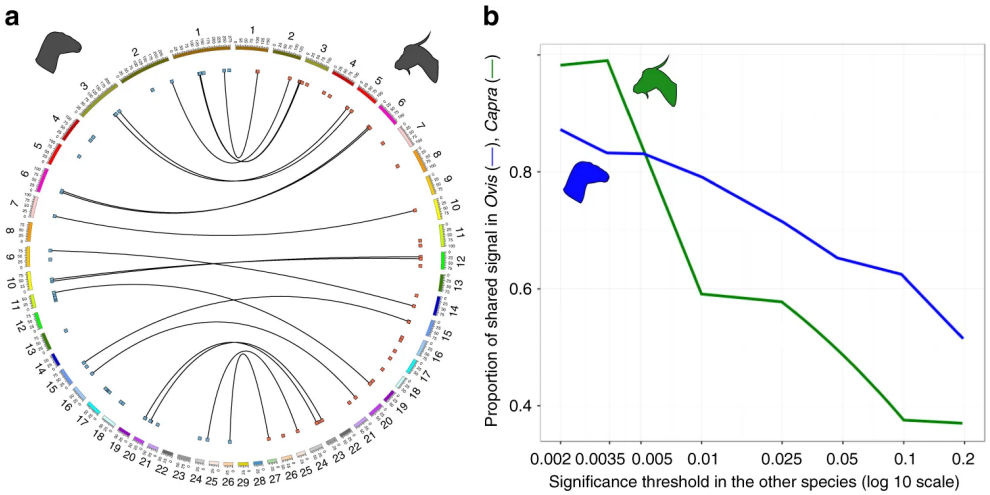


Figure 3.10 : Signatures de domestication chez les chèvres et les moutons. (a) Représentation des régions significatives sur le génomes de la chèvre (droite) et du mouton (gauche) (b) Convergence dans le signal de sélection entre les deux espèces. Tirée de Alberto *et al.* (2018)

En se basant sur les résultats obtenus dans les études listées dans le tableau 3.1, et plus particulièrement sur celles appliquées aux ruminants (chèvre, mouton et vache) il apparaît que certaines régions génomiques, voire certains gènes candidats, ont été des cibles récurrentes de la sélection. Il s'agit d'une part de gènes connus pour leur implication dans le déterminisme de la couleur de la robe ou de la peau des animaux : KIT, ASIP, MC1R en particulier. Ceci peut s'expliquer par l'importance de l'aspect extérieur des animaux dans la définition des races associée à un relâchement de la pression de sélection sur des phénotypes non observés chez les espèces sauvages.

2. *e.g.* Figure S16 dans Boitard *et al.* (2016)
3. cf. Figure 3 de Rochus *et al.* (2018)
4. cf. Figure 4 ibid.

Certains de ces gènes sont par ailleurs des régulateurs majeurs du développement des cellules de la crête neurale. Les mutations affectant leur régulation peuvent donc avoir des effets pléiotropes sur d'autres caractères, comme le suggère l'hypothèse postulant l'importance des cellules de la crête neurale dans le syndrome de domestication (Wilkins *et al.*, 2014; Sánchez-Villagra Marcelo R. *et al.*, 2016). D'autres gènes associés à des signatures de sélection dans des espèces différentes sont associés à la morphologie et à la croissance. Il s'agit par exemple du gène codant pour la myostatine (GDF8 / MSTN) une protéine régulatrice de la croissance musculaire ou le locus associant les gènes NCAPG et LCORL, en fort déséquilibre de liaison et associés à des caractères de croissance chez de nombreuses espèces de mammifères. L'exemple de ces gènes sont extrêmes mais il est globalement fréquent de retrouver les mêmes régions génomiques associées à des signatures de sélection dans des espèces différentes. Dans l'étude de **Alberto *et al.* (2018)**, nous avons cherché à quantifier dans quelle mesure ces co-localisations des signatures de domestication chez la chèvre et le mouton (Figure 3.10.a) étaient dues au hasard. Nous avons appliqué une approche de FDR stratifié (Sun *et al.*, 2006) en estimant la part de SNP significatifs dans une espèce dans les régions détectées dans l'autre (Figure 3.10.b). Cette approche montre clairement une convergence non aléatoire des signatures de sélection dans les deux espèces.

Si les caractères liés à la pigmentation et à la morphologie des animaux semblent clairement avoir été objet de processus adaptatifs, les études menées sur des panels de populations mondiales semblent également révéler des processus adaptatifs liés à l'adaptation environnemental des animaux. Il s'agit d'une part d'adaptation climatique, comme par exemple pour le gène TRPM8 chez le mouton lié à la perception du froid et récemment identifié comme signature de sélection chez les humains (Key *et al.*, 2018), ou des gènes de développement tel que le cluster HOXC et son interacteur PBX1 associé à une adaptation à des régions de fortes températures annuelles. Enfin, chez les petits ruminants ont été identifié plusieurs gènes impliqué dans l'horloge circadienne (SOX14, NOCT, RAI, TH, TSHR) pouvant affecter la saisonalité de la reproduction.

3.5 Conclusion

Mes recherches sur les méthodes de détection de signatures adaptatives m'ont conduit à participer au développement de statistiques particulièrement adaptées aux jeux de données produits dans le cadre de l'étude de la diversités des animaux domestiques. En plus du développement de ces statistiques, nous avons mis au point des procédures permettant d'interpréter finement les signatures obtenues pour identifier les populations sélectionnées et les allèles ou haplotypes adaptatifs. Ces développements sont distribués à la communauté sous la forme d'un logiciel et de scripts associés permettant de reproduire les résultats obtenus sur des jeux de données réels. La possibilité qui m'a été offerte de travailler sur des jeux de données de grande dimension m'a permis d'améliorer les performances du logiciel et de progresser dans la compréhension des processus adaptatifs des populations animales. Je compte poursuivre ces développements en particulier pour mieux modéliser les processus adaptatifs et passer de tests basés sur l'ajustement d'un modèle neutre à des inférences quantitatives de paramètres adaptatifs, comme je vais le décrire dans le chapitre 4.

4. Projet de recherche : évolution des caractères complexes dans les populations animales

Les populations animales domestiques offrent de bons modèles pour caractériser les processus adaptatifs. Elles disposent de ressources génomiques de grande ampleur, sont caractérisées phénotypiquement et leur histoire est documentée. Mon projet de recherche s'articule autour de l'étude de ces populations en utilisant des jeux de données de plus en plus importants et de structure variées. Dans cette partie de mon rapport je vais tout d'abord présenter les questions méthodologiques que je compte aborder puis deux caractères quantitatifs que je compte plus spécifiquement étudier.

4.1 Modélisation des effets adaptatifs

Le principe des modèles sur lesquels j'ai travaillé jusqu'à présent est basé sur le test d'ajustement d'un modèle neutre. S'il est possible d'annoter dans une certaine mesure les évènements adaptatifs à partir des déviations au modèle neutre (cf. chapitre 3), cette approche est limitée dans le fait qu'elle ne permet pas d'estimer des modèles explicites d'adaptation. Par ailleurs ce sont des approches qui testent individuellement les locus du génome. Je travaille actuellement sur des modèles qui sont basés sur une spécification des hypothèses alternatives et que je voudrais développer pour modéliser conjointement les effets pan-génomiques, à la manière des modèles utilisés pour les études d'association ou la sélection génomique. Ce travail de modélisation s'effectue dans deux contextes différents : l'analyse de données sur des populations structurées (type FLK, hapFLK) et l'analyse de données sur des échantillons répartis dans le temps.

Modèles de différentiation

Le principe d'extension du modèle FLK pour inclure des effets adaptatifs a été proposé par Coop *et al.* (2010) puis amélioré par Günther et Coop (2013) et Gautier (2015). Si l'on considère la formulation du modèle FLK sous la forme d'un modèle linéaire (équation (3.2)), son extension proposée par les références ci-dessus est naturelle :

$$\boldsymbol{p} \sim \mathcal{N}(p_0\mathbf{1} + \mathbf{x}\boldsymbol{\beta}, p_0(1-p_0)\mathbf{F}) \tag{4.1}$$

ici $\boldsymbol{x}$ représente un vecteur (potentiellement une matrice) de covariable(s) observée(s) (typiquement des mesures environnementales) sur les populations et $\boldsymbol{\beta}$ les effets associés à ces covariables sur l'évolution des fréquences alléliques. Il s'agit donc bien d'une approche qui n'est plus basée sur le rejet du modèle neutre mais sur une spécification des effets adaptatifs $\boldsymbol{\beta}$. En revanche dans l'application qui est faite de ces modèles, la quantité d'intérêt utilisée est généralement une

statistique de test du modèle (4.1) vs. le modèle (3.2), pas les valeurs des effets $\boldsymbol{\beta}$. Ceci peut se justifier par le fait que dans le cas où $\boldsymbol{x}$ est une variable environnementale ces valeurs ne sont pas particulièrement interprétables au delà de leurs signes.

Dans le cadre de la thèse de Jason Lapeyronnie, nous avons étudié les extensions de FLK et hapFLK basées sur le modèle (4.1). Nous avons développé deux types de statistiques de tests, l'une utilisant une approche par maximum de vraisemblance et l'autre une approche Bayésienne comme les références ci-dessus mais en utilisant les techniques que j'ai développé en post-doctorat (cf. chapitre 2) qui permettent de s'affranchir de l'utilisation d'algorithmes MCMC et sont donc beaucoup plus rapides. Sur la base de ces développements, je voudrais maintenant m'intéresser à la modélisation des effets $\boldsymbol{\beta}$ d'une part en cherchant à leur donner une interprétation adaptative (*e.g.* en terme de coefficient de sélection) et d'autre part en considérant un modèle hiérarchique qui permette de relier les effets adaptatifs de tous les locus entre eux.

Analyse de données génétiques temporelles

Le développement des techniques de séquençage et la baisse des coûts permettent aujourd'hui de disposer de données génétiques temporelles dans plusieurs contextes. Les techniques de séquençage d'ADN issu de fossiles a vu naître une nouvelle discipline : la paléo-génomique. Je suis partenaire d'un projet ANR destiné à étudier l'histoire des populations bovines depuis leur domestication jusqu'à aujourd'hui. Dans ce projet, nous aurons accès à quelques dizaines de séquences de fossiles répartis dans le temps. Un autre contexte dans lequel apparaissent des séries temporelles est celui de populations de sélection expérimentale. Au département de génétique animale, cette approche est largement utilisée pour étudier le déterminisme génétique des caractères à travers la réponse à leur sélection. Dans ces populations, des échantillons d'ADN sont conservés au cours des générations et la baisse des coûts permet aujourd'hui de les exploiter. Enfin, le développement de la sélection génomique produit chaque année de nouveaux génotypages. Ceci conduit à la création en continu de cohortes d'individus suivies dans le temps dans les espèces et populations qui implémentent cette méthode de sélection.

Dans ce contexte nous avons, dans le cadre de la thèse de Cyriel Paris, commencé à développer une méthode d'estimation des coefficients de sélection à partir de données génétiques temporelles **(Paris *et al.*, 2019)**. Le principe de modélisation que nous avons utilisé est un modèle de Markov caché (Figure 4.1). Les observations Y_k sont les comptages alléliques observés à un locus à la date t_k (en générations). Elles considérées comme des échantillons aléatoires tirés dans une loi binomiale (g_k) dont le paramètre de succès est la fréquence allélique dans la population à cette date, notée X_k. Les changements de fréquences alléliques au cours du temps sont déterminés par les lois de transition $P(X_k|X_{k-1}) \sim Q_k(N, s)$. Elles dépendent d'une part de la taille de la population (N) et d'autre part du coefficient de sélection de l'allèle (s). Notre travail a consisté à étudier différents modèles de transition : le modèle génétique de Wright-Fisher avec sélection est justifié biologiquement mais conduit à des problèmes de temps de calcul, nous avons donc étudié différentes approximations de ce modèle pour retenir celui proposé par Tataru *et al.* (2015) basé sur une loi Beta associée à des probabilités d'absorption en 0 et 1 (modèle Beta-à-pics). Nous avons pu montrer que ce modèle permettait d'obtenir les mêmes résultas d'inférence statistique que le modèle de Wright-Fisher tout en étant plus performant d'un point de vue calculatoire. Ce travail

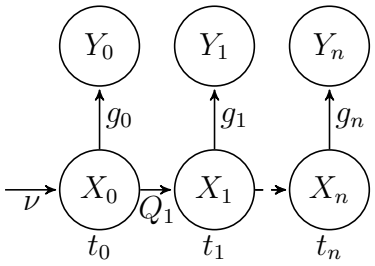


FIGURE 4.1 : Modèle de Markov Caché pour des données génétiques temporelles.

nous fournit aujourd'hui une méthode pour calculer efficacement la vraisemblance de paramètres démographiques sur la base de données génétiques temporelles (*i.e.* $P(\boldsymbol{Y}|N, s)$). Nous voudrions maintenant exploiter ce modèle en considérant l'ensemble des locus du génome. Ceci pour (i) estimer conjointement N et s (actuellement sur des données réelles N est estimé indépendamment de s), (ii) autoriser des changements de taille de population au cours du temps et (iii) modéliser conjointement les effets adaptatifs des locus (s) avec un modèle hiérarchique comme pour l'approche précédente.

Applications

Je disposerais de plusieurs jeux de données qui seront utiles pour le développement, la validation et l'évaluation des performances de ces modèles. Pour la première approche, dans le cadre du projet européen Treasure, nous avons accès au séquençage en mélange de vingts populations de populations locales de porcs pour lesquels nous avons également des mesures phénotypiques. Dans le cadre de la thèse de Klavdija Poklukar, nous avons commencé à analyser ce jeu de données. Nous allons également travaillé dans le cadre du projet européen Smarter sur les données du projet VarGoats [1] qui a produit des séquences individuelles d'environ 900 chèvres de populations mondiales. En ce qui concerne les données temporelles, nous avons déjà appliqué la méthode sur des lignées divergentes de poulet **(Paris *et al.*, 2019)** et nous nous en servirons dans le cadre du projet ANR Path2Bos sur des données d'ADN de fossiles.

A moyen terme, j'aimerai utiliser nos annotations sur les effets adaptatifs du génome pour les combiner aux annotations fonctionnelles basées sur des données moléculaires. Ces annotations sont produites dans le cadre de l'initiative internationale FAANG [2]. Mon équipe participe à la contribution de l'INRA à cette initiative à travers le travail de Sylvain Foissac, David Robelin et de Sarah Djebali-Quelen (Foissac *et al.*, 2019, www.fragencode.org). L'objectif est de permettre d'affiner l'annotation des régions régulatrices sur la base d'une prédiction de leur importance évolutive. Le principe de cette intégration pourrait se baser sur la méthode fitCons (Gulko *et al.*, 2015). Le principe de cette méthode est tout d'abord de segmenter le génome sur la base des patrons d'annotation fonctionnelle (624 classes de segments) pour ensuite estimer au sein de chaque classe la probabilité qu'une mutation ait un impact sur la valeur sélective. Les modèles décrit ci-dessus devraient permettre d'estimer ces probabilités. Cela demandera par ailleurs d'adapter le modèle de

1. `www.goatgenome.org/vargoats.html`
2. `www.faang.org`

segmentation aux données générées par FAANG qui ne sont pas exactement les mêmes que celles du projet ENCODE humain et nécessitera donc une collaboration étroite entre bioinformaticiens et généticiens des populations.

4.2 Evolution du processus de recombinaison

Récemment, je me suis intéressé plus spécifiquement au processus biologique fondamental qu'est la recombinaison méiotique. L'étude de ce processus biologique a d'abord des objectifs académiques mais l'établissement de cartes génétiques précises est également fondamental pour l'ensemble des études de génétiques comme par exemple pour optimiser des schémas de sélection (Zhong et Jannink, 2007, thèse Alice Danguy-des-Déserts) ou la détection de QTL **(Petit *et al.*, 2018b)**. L'étude de la recombinaison ovine que nous avons mené dans le cadre de la thèse de Morgane Petit est un bon exemple d'analyse issue de la science des données **(Petit *et al.*, 2017)** : nous avons utilisé des données produites dans un contexte différent (la sélection génomique en race Lacaune) en les combinant avec des jeux de données existants pour étudier ce même phénomène (Johnston *et al.*, 2016) ou pour des études de diversité **(Rochus *et al.*, 2018)**. Dans le cadre de mon projet de recherche, je voudrais explorer plus avant ce processus biologique en m'intéressant plus spécifiquement à son évolution, sous différents aspects :

1. Nos travaux précédents ont montré une très bonne conservation des cartes génétiques entre populations Soay et Lacaune. J'aimerai confirmer cette conservation à court terme en établissant des cartes de recombinaison dans d'autres populations ovines. Dans un premier temps, je pense m'appuyer sur la population Manech qui démarre un programme de sélection génomique.
2. Je voudrais étudier la variabilité de la localisation de la recombinaison sur le génome. Ce type de phénotype a un déterminisme génétique différent de celui gouvernant l'intensité de recombinaison et semble avoir une évolution rapide. Chez l'homme et la souris un gène majeur (PRDM9) a été identifié comme impliqué dans ce déterminisme. De fortes présomptions sur son action chez le bovin existent (Sandor *et al.*, 2012 ; Ma *et al.*, 2015). Je dispose de données dans la race Romane qui devraient permettre d'étudier cette question chez les ovins.
3. Si les cartes génétiques sont conservées entre populations d'une même espèce, j'aimerai identifier leur évolution potentielle sur de plus grandes échelles de temps en utilisant une approche trans-spécifique. En effet, nous avons la chance de disposer de ressources similaires chez trois espèces de ruminants proches phylogénétiquement : la vache, le mouton et la chèvre. Les données sont disponibles chez la vache (Kadri *et al.*, 2016) et le mouton et nous avons commencé avec Rachel Rupp de l'équipe *Génétique et Sélection des Petits Ruminants* du laboratoire à établir ces cartes chez la chèvre. A partir de ces données, je voudrais établir une comparaison analytique des cartes de recombinaison chez les trois espèces et identifier les facteurs affectant la conservation (ou son absence) des patrons de recombinaison entre espèces.

Ce projet demande d'établir des modèles statistiques sur la variabilité de la recombinaison entre groupes d'observations, correspondant soit à des populations différentes soit à des espèces différentes. Dans le cadre de la thèse d'Alice Danguy-des-Déserts, qui travaille sur le blé tendre, nous avons commencé à mettre au point ces modèles. Il faudra ensuite les adapter pour travailler sur les espèces animales et effectuer l'analyse des données disponibles. Au delà de cette approche génétique, d'autres membres du laboratoire s'intéressent à la recombinaison par des approches différentes : Alain Pinton et l'équipe *Cytogénomique structurale et fonctionnelle* (Cytogen) s'y intéresse à partir d'approches cytogénétiques et nous avons discuté avec Hervé Acloque du laboratoire GABI de la possibilité d'utiliser des approches moléculaires (Pratto *et al.*, 2014 ; Lange *et al.*, 2016). Ces approches pourraient permettre d'établir des validations expérimentales de nos résultats statistiques.

4.3 Résistance des abeilles mellifères à *Varroa destructor*

Ces dernières années ont vu l'inquiétude augmenter autour du déclin des populations d'abeilles mellifères (*Apis Mellifera*) en France et dans le monde. Les raisons de ce déclin sont multi-factorielles et l'un des facteurs soupçonné de contribuer de manière importante à ce déclin est l'apparition d'un acarien parasite de l'abeille : *Varroa destructor*. Ce parasite est originaire d'Asie du Sud-Est et est apparu en France au début des années 1980. Je suis impliqué dans l'étude du déterminisme génétique de la résistance des abeilles à ce parasite dans le cadre d'un projet financé par le ministère de l'agriculture nommé BeeStrong. Ce projet a pour but final d'effectuer des études d'association des polymorphismes des populations françaises d'abeille avec des phénotypes de résistance à Varroa et d'identifier des marqueurs génétiques utilisables pour la gestion des populations et la sélection. L'échantillonage et le phénotypage des colonies d'abeilles est géré par nos collaborateurs de l'INRA d'Avignon (Fanny Mondet) et l'Institut Technique et Scientifique au Service de l'Apiculture (ITSAP), la gestion des échantillons d'ADN par l'entreprise Labogena et l'INRA est responsable de la mise au point de programmes de sélection (GABI Jouy-en-Josas, Florence Phocas) et du séquençage des colonies (GenPhySE, équipe Cytogen, Alain Vignal). Je suis impliqué dans le développement de modèles statistiques pour analyser ces données dans le cadre du post-doctorat de Sonia Eynard.

La première étape de nos travaux pour conduire les études d'association chez l'abeille a été de résoudre le problème du génotypage des individus. Travailler sur cette espèce nécessite de développer de nouveaux modèles statistiques, les approches existantes n'étant pas adaptées à ce type d'organismes. En effet, chez l'abeille mellifère, une grande proportion des phénotypes, en particulier ceux liés à la résistance à Varroa, sont définis au niveau de la colonie. D'un point de vue génétique, une colonie est fondée par une reine diploïde initialement fécondée par plusieurs mâles. Un mâle est issu d'un gamète de reine non fécondé et est donc parfaitement haploïde. Une colonie est donc constituée d'abeilles ouvrières qui sont issues d'une même reine mais qui peuvent avoir comme père des mâles différents. Ainsi les structures d'apparentement entre ouvrières d'une même colonie sont complexes car elles contiennent des paires d'ouvrières ayant reçu le même gamète mâle (dites "super-sœurs"), des pleines sœurs et des demi-sœurs. Pour les études d'association, on peut considérer le "génotype" d'une colonie de différentes façons. Il peut s'agir de la fréquence des al-

lèles au sein du groupe d'ouvrières, cependant ce "génotype" n'est pas transmis aux reines filles qui fonderont de nouvelles colonies ou aux mâles qui féconderont de nouvelles reines. On peut aussi le caractériser en faisant abstraction de la contribution des mâles et le résumer au génotype de la reine qui, lui, est transmis de façon Mendélienne que ce soit aux reines filles ou aux mâles. La difficulté vient alors du fait qu'il n'est pas possible de génotyper la reine sans la sacrifier. Une solution possible est de génotyper des mâles de la colonie ce qui revient à génotyper des gamètes de la reine. Cette approche a deux inconvénients : le premier est qu'il faut effectuer plusieurs génotypages par reine ce qui implique des coûts élevés, le deuxième est que nous n'avons dans ce cas aucune information sur la contribution mâle à la génétique de la colonie. Pour ces raisons, nous avons employé une autre approche qui consiste à génotyper un mélange de nombreuses ouvrières (≈ 500). Dans ce mélange pour chaque polymorphisme, 50% des allèles proviennent de la reine et 50% du lot de mâles qui l'ont fécondée. La reconstruction du génotype de la reine à partir de ses données demande alors de développer des modèles statistiques spécifiques.

Dans le cadre du projet Beestrong, 1500 colonies d'abeilles françaises sont en cours de séquençage. Nous disposons aujourd'hui d'une partie de ces données, comprenant environ 5 millions de marqueurs pour près de 1000 colonies, et nous avons commencé à développer différents modèles statistiques pour reconstruire le génotype des reines. Ces modèles sont basés sur la vraisemblance :

$$P(x_h, d_h | f_h, g_h) = \binom{d_h}{x_h} \left(\frac{f_h + g_h}{2}\right)^{x_h} \left(1 - \frac{f_h + g_h}{2}\right)^{d_h - x_h} \tag{4.2}$$

où h indice la colonie, d_h est le nombre d'allèles observés au locus considéré (*i.e.* la profondeur de séquençage), x_h est le nombre d'allèle de référence parmi eux, f_h est la fréquence de l'allèle de référence parmi les mâles ayant fécondé la reine et g_h est le génotype de la reine exprimé en fréquence (0, 1/2 ou 1). Sur la base de cette vraisemblance, nous étudions plusieurs approches. La première consiste à considérer que la population de colonies est une population panmictique homogène. Dans ce cas f_h est un paramètre partagé par toutes les colonies ($\forall h f_h = f$) et la vraisemblance (4.2) peut être exprimée en fonction de f uniquement :

$$\mathcal{L}_h(f) = P(x_h, d_h | f) = \sum_{g_h} P(x_h, d_h | f, g_h) P(g_h | f) \tag{4.3}$$

où $P(g_h|f)$ dépend uniquement de f. f peut être estimé en maximisant la vraisemblance combinant toutes les colonies ($\mathcal{L}(f) = \prod_h \mathcal{L}_h(f)$), et on peut prédire le génotype de la reine par $P(g_h|\hat{f}, x_h, d_h)$. L'hypothèse la plus problématique dans cette modélisation est celle d'une population homogène non structurée que nous savons être fausse. Nous nous tournons à présent vers des approches permettant d'intégrer la possibilité de structuration en groupes correspondant aux différentes sous espèces répertoriées (abeille noire, abeille jaune, abeille caucasienne), en nous basant sur des modèles bien connus de la structure génétique des populations (Pritchard *et al.*, 2000 ; Falush *et al.*, 2003). Par la suite, nous validerons nos résultats sur certaines colonies pour lesquelles des mâles ont été séquencés et effectuerons des études d'association avec différents caractères liés à la résistance à Varroa.

Ces données seront également utiles pour étudier l'histoire des populations d'abeilles, en particulier des différents types génétiques qui la compose. Ceci nécessitera de travailler à partir de mo-

dèles individu centré (Kelleher *et al.*, 2018 ; Albers et McVean, 2019 ; Speidel *et al.*, 2019 ; Duforet-Frebourg *et al.*, 2014, 2016) car, contrairement aux espèces animales sur lesquelles nous avons l'habitude de travailler, la structure génétique est continue beaucoup d'individus étant issus de croisement entre types génétiques différents. Il sera également intéressant d'étudier la structure spatiale de la diversité génétique de cette population (Petkova *et al.*, 2016).

4.4 Conclusion Générale

L'augmentation de la quantité de données disponibles pour les études de génétique des populations animales offre de grandes opportunités pour mieux comprendre les effets génétiques contribuant au déterminisme des caractères et leur dynamique évolutive (Sella et Barton, 2019). Les espèces animales domestiques sont des objets d'étude qui ont à la fois des avantages pour étudier cette question mais également certains inconvénients. Les avantages proviennent de l'histoire évolutive de ces populations que l'on peut considérer comme des populations expérimentales à grande échelle, de part le fait qu'elles ont connu et continuent de connaître des processus évolutifs importants sur des échelles de temps courts. Leur temps de génération est relativement court (quelques années) ce qui permet d'avoir accès à des temps d'évolution relativement long rapportés au temps calendaire. Par ailleurs, ces processus ont été répétés dans plusieurs populations mais également chez plusieurs espèces proches phylogénétiquement. L'analyse comparative de ces espèces est donc particulièrement intéressante. L'étude de ces populations est par ailleurs facilitée par la disponibilité d'outils et de ressources génomiques importants du fait de leur intérêt socio-économique. Cet importance est aussi source de pluridisciplinarité car il est assez typique que les mêmes populations voire les mêmes individus soient étudiés pour des questions différentes. Au rang des inconvénients se situent la forte structuration au sein des populations et la forte dérive qui peut masquer les effets adaptatifs. Cette structuration rend aussi compliquée l'estimation individuelle des effets des polymorphismes car le déséquilibre de liaison s'étend sur de longues distances. De récents résultats sur l'étude de l'adaptation polygénique dans les populations humaines européennes (Berg *et al.*, 2019) pourtant beaucoup moins structurées incitent à réfléchir à des modèles statistiques et/ou des plans d'expériences spécifiques pour étudier ces questions dans les populations animales.

D'un point de vue méthodologique, mon projet de recherche fait appel dans différents contextes à l'utilisation de modèles hiérarchiques. De récents développements dans le domaine de la statistique Bayésienne empirique seront certainement des outils à prendre en main pour modéliser des collections d'effets adaptatifs ou génétiques de grande dimension. Je pense en particulier aux méthodes de shrinkage adaptatif (Stephens, 2017 ; Wang, 2017) qui permettent de modéliser de manière très flexible les distributions empiriques d'effets estimés comme mélanges de lois variées. Je pense également évaluer l'intérêt de nouveaux outils méthodologiques qui se développent actuellement dans le domaine de l'intelligence artificielle. Il s'agit par exemple des méthodes de prédiction et d'inférence utilisant des réseaux de neurones profonds ou des forêts aléatoires qui commencent à être appliquées dans le domaine de la génétique des populations (Sheehan et Song, 2016 ; Flagel *et al.*, 2019 ; Raynal *et al.*, 2019) ou la détection de polymorphismes à partir de données de reséquençage (Poplin *et al.*, 2018).

Bibliographie

Albers, P. K. et McVean, G. (2019). Dating genomic variants and shared ancestry in population-scale sequencing data. *bioRxiv*, page 416610. (cf. page 57).

Alberto, F. J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., de Villemereuil, P., Benjelloun, B., Librado, P., Biscarini, F., Colli, L., Barbato, M., Zamani, W., Alberti, A., Engelen, S., Stella, A., Joost, S., Ajmone-Marsan, P., Negrini, R., Orlando, L., Rezaei, H. R., Naderi, S., Clarke, L., Flicek, P., Wincker, P., Coissac, E., Kijas, J., Tosser-Klopp, G., Chikhi, A., Bruford, M. W., Taberlet, P. et Pompanon, F. (2018). Convergent genomic signatures of domestication in sheep and goats. *Nat Commun*, 9 :813. (cf. pages 44, 45, 47, 48, and 49).

Almasy, L. et Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet*, 62(5) :1198–211. (cf. page 24).

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., Boyle, E. A., Zhang, X., Racimo, F., Pritchard, J. K. et Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. *eLife*, 8 :e39725. (cf. page 57).

Bertolini, F., Servin, B., Talenti, A., Rochat, E., Kim, E. S., Oget, C., Palhière, I., Crisà, A., Catillo, G., Steri, R., Amills, M., Colli, L., Marras, G., Milanesi, M., Nicolazzi, E., Rosen, B. D., Van Tassell, C. P., Guldbrandtsen, B., Sonstegard, T. S., Tosser-Klopp, G., Stella, A., Rothschild, M. F., Joost, S. et Crepaldi, P. (2018). Signatures of selection and environmental adaptation across the goat genome post-domestication. *Genet Sel Evol*, 50 :57. (cf. pages 44 and 45).

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisà, A., de León, F. A. P., Schwartz, J. C., Hammond, J. A., Waldbieser, G. C., Schroeder, S. G., Liu, G. E., Dunham, M. J., Shendure, J., Sonstegard, T. S., Phillippy, A. M., Tassell, C. P. V. et Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, 49(4) :643. (cf. page 10).

Boitard, S., Boussaha, M., Capitan, A., Rocha, D. et Servin, B. (2016). Uncovering adaptation from sequence data : Lessons from genome resequencing of four cattle breeds. *Genetics*, 203(1) :433–450. (cf. pages 43, 45, 46, 47, and 48).

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S. et Sancristobal, M. (2010). Detecting selection in population trees : the Lewontin and Krakauer test extended. *Genetics*, 186(1) :241–62. (cf. pages 38 and 41).

Bouwman, A. C., Daetwyler, H. D., Chamberlain, A. J., Ponce, C. H., Sargolzaei, M., Schenkel, F. S., Sahana, G., Govignon-Gion, A., Boitard, S., Dolezal, M., Pausch, H.,

Brøndum, R. F., Bowman, P. J., Thomsen, B., Guldbrandtsen, B., Lund, M. S., Servin, B., Garrick, D. J., Reecy, J., Vilkki, J., Bagnato, A., Wang, M., Hoff, J. L., Schnabel, R. D., Taylor, J. F., Vinkhuyzen, A. A. E., Panitz, F., Bendixen, C., Holm, L.-E., Gredler, B., Hozé, C., Boussaha, M., Sanchez, M.-P., Rocha, D., Capitan, A., Tribout, T., Barbat, A., Croiseau, P., Drögemüller, C., Jagannathan, V., Vander Jagt, C., Crowley, J. J., Bieber, A., Purfield, D. C., Berry, D. P., Emmerling, R., Götz, K.-U., Frischknecht, M., Russ, I., Sölkner, J., Van Tassell, C. P., Fries, R., Stothard, P., Veerkamp, R. F., Boichard, D., Goddard, M. E. et Hayes, B. J. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet*, 50 :362–367. (cf. pages 44, 45, and 47).

Boyle, E. A., Li, Y. I. et Pritchard, J. K. (2017). An Expanded View of Complex Traits : From Polygenic to Omnigenic. *Cell*, 169(7) :1177–1186. (cf. page 33).

Churchill, G. A., Airey, D. C., Allayee, H., Angel, J. M., Attie, A. D., Beatty, J., Beavis, W. D., Belknap, J. K., Bennett, B., Berrettini, W., Bleich, A., Bogue, M., Broman, K. W., Buck, K. J., Buckler, E., Burmeister, M., Chesler, E. J., Cheverud, J. M., Clapcote, S., Cook, M. N., Cox, R. D., Crabbe, J. C., Crusio, W. E., Darvasi, A., Deschepper, C. F., Doerge, R. W., Farber, C. R., Forejt, J., Gaile, D., Garlow, S. J., Geiger, H., Gershenfeld, H., Gordon, T., Gu, J., Gu, W., de Haan, G., Hayes, N. L., Heller, C., Himmelbauer, H., Hitzemann, R., Hunter, K., Hsu, H. C., Iraqi, F. A., Ivandic, B., Jacob, H. J., Jansen, R. C., Jepsen, K. J., Johnson, D. K., Johnson, T. E., Kempermann, G., Kendziorski, C., Kotb, M., Kooy, R. F., Llamas, B., Lammert, F., Lassalle, J. M., Lowenstein, P. R., Lu, L., Lusis, A., Manly, K. F., Marcucio, R., Matthews, D., Medrano, J. F., Miller, D. R., Mittleman, G., Mock, B. A., Mogil, J. S., Montagutelli, X., Morahan, G., Morris, D. G., Mott, R., Nadeau, J. H., Nagase, H., Nowakowski, R. S., O'Hara, B. F., Osadchuk, A. V., Page, G. P., Paigen, B., Paigen, K., Palmer, A. A., Pan, H. J., Peltonen-Palotie, L., Peirce, J., Pomp, D., Pravenec, M., Prows, D. R., Qi, Z., Reeves, R. H., Roder, J., Rosen, G. D., Schadt, E. E., Schalkwyk, L. C., Seltzer, Z., Shimomura, K., Shou, S., Sillanpaa, M. J., Siracusa, L. D., Snoeck, H. W., Spearow, J. L., Svenson, K., Tarantino, L. M., Threadgill, D., Toth, L. A., Valdar, W., de Villena, F. P., Warden, C., Whatley, S., Williams, R. W., Wiltshire, T., Yi, N., Zhang, D., Zhang, M. et Zou, F. (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet*, 36(11) :1133–7. (cf. page 24).

Coop, G., Wen, X., Ober, C., Pritchard, J. K. et Przeworski, M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319(5868) :1395–8. (cf. page 13).

Coop, G., Witonsky, D., Di Rienzo, A. et Pritchard, J. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4) :1411–1423. (cf. page 51).

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A. et Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics*, 36(7) :700–706. (cf. page 2).

Crepieux, S., Lebreton, C., Servin, B. et Charmet, G. (2004). Quantitative trait loci (QTL) detection in multicross inbred designs : recovering QTL identical-by-descent status information from marker data. *Genetics*, 168(3) :1737–49. (cf. pages 24 and 25).

Delaneau, O., Marchini, J. et Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2) :179–181. (cf. page 1).

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., Tosser-Klopp, G., Wang, J., Yang, S., Liang, J., Chen, W., Chen, J., Zeng, P., Hou, Y., Bian, C., Pan, S., Li, Y., Liu, X., Wang, W., Servin, B., Sayre, B., Zhu, B., Sweeney, D., Moore, R., Nie, W., Shen, Y., Zhao, R., Zhang, G., Li, J., Faraut, T., Womack, J., Zhang, Y., Kijas, J., Cockett, N., Xu, X., Zhao, S., Wang, J. et Wang, W. (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat Biotechnol*, 31(2) :135–41. (cf. page 9).

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(3) :789–98. (cf. page 13).

Druet, T. et Georges, M. (2015). LINKPHASE3 : an improved pedigree-based phasing algorithm robust to genotyping and map errors. *Bioinformatics*, 31(10) :1677–9. (cf. pages 13 and 16).

Du, X., Servin, B., Womack, J. E., Cao, J., Yu, M., Dong, Y., Wang, W. et Zhao, S. (2014). An update of the goat genome assembly using dense radiation hybrid maps allows detailed analysis of evolutionary rearrangements in Bovidae. *BMC Genomics*, 15 :625. (cf. page 9).

Duforet-Frebourg, N., Bazin, E. et Blum, M. G. (2014). Genome scans for detecting footprints of local adaptation using bayesian factor model. *Molecular Biology of Evolution*, 31(9) :2483–2495. (cf. page 57).

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E. et Blum, M. G. (2016). Detecting Genomic Signatures of Natural Selection with Principal Component Analysis : Application to the 1000 Genomes Data. *Mol Biol Evol*, 33(4) :1082–93. (cf. page 57).

Falush, D., Stephens, M. et Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data : Linked Loci and Correlated Allele Frequencies. *Genetics*, 164(4) : 1567–1587. (cf. page 56).

Faraut, T., de Givry, S., Chabrier, P., Derrien, T., Galibert, F., Hitte, C. et Schiex, T. (2007). A comparative genome approach to marker ordering. *Bioinformatics*, 23(2) :e50–6. (cf. pages 4 and 5).

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M. et Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured population. *Genetics*, 193 :929–941. (cf. pages 42 and 43).

Fariello, M.-I., Servin, B., Tosser-Klopp, G., Rupp, R., Moreno, C., Consortium, I. S. G., Cristobal, M. S. et Boitard, S. (2014). Selection signatures in worldwide sheep populations. *PLOS ONE*, 9(8) :1–12. (cf. pages 43, 44, 45, and 46).

Favier, A. (2011). *Décompositions fonctionnelles et structurelles dans les modèles graphiques probabilistes appliquées à la reconstruction d'haplotypes*. Thèse de doctorat, Université Paul Sabatier. (cf. page 13).

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2) :399–433. (cf. page 24).

Flagel, L., Brandvain, Y. et Schrider, D. R. (2019). The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution*, 36(2) :220–238. (cf. page 57).

Fledel-Alon, A., Leffler, E. M., Guan, Y., Stephens, M., Coop, G. et Przeworski, M. (2011). Variation in human recombination rates and its genetic determinants. *PLoS One*, 6(6) : e20321. (cf. page 13).

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., Esquerré, D., Zytnicki, M., Derrien, T., Bardou, P., Blanc, F., Cabau, C., Crisci, E., Dhorne-Pollet, S., Drouet, F., Faraut, T., Gonzalez, I., Goubil, A., Lacroix-Lamandé, S., Laurent, F., Marthey, S., Marti-Marimon, M., Momal-Leisenring, R., Mompart, F., Quéré, P., Robelin, D., Cristobal, M. S., Tosser-Klopp, G., Vincent-Naulleau, S., Fabre, S., der Laan, M.-H. P.-V., Klopp, C., Tixier-Boichard, M., Acloque, H., Lagarrigue, S. et Giuffra, E. (2019). Transcriptome and chromatin structure annotation of liver, CD4+ and CD8+ T cells from four livestock species. *bioRxiv*, page 316091. (cf. page 53).

Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, 201(4) :1555–1579. (cf. page 51).

George, A. W., Visscher, P. M. et Haley, C. S. (2000). Mapping quantitative trait loci in complex pedigrees : a two-step variance component approach. *Genetics*, 156(4) :2081–92. (cf. page 24).

Gholami, M., Reimer, C., Erbe, M., Preisinger, R., Weigend, A., Weigend, S., Servin, B. et Simianer, H. (2015). Genome scan for selection in structured layer chicken populations exploiting linkage disequilibrium information. *PLOS ONE*, 10(7) :1–15. (cf. page 45).

Good, I. (1992). The Bayes/non-Bayes compromise : a brief review. *Journal of the American Statistical Association*, 87 :597–606. (cf. page 30).

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H. J., Li, S., Larkin, D. M., Kim, H., Frantz, L. A., Caccamo, M., Ahn, H., Aken, B. L., Anselmo, A., Anthon, C., Auvil, L., Badaoui, B., Beattie, C. W., Bendixen, C., Berman, D., Blecha, F., Blomberg, J., Bolund, L., Bosse, M., Botti, S., Bujie, Z., Bystrom, M., Capitanu, B., Carvalho-Silva, D., Chardon, P., Chen, C., Cheng, R., Choi, S. H., Chow, W., Clark, R. C., Clee, C., Crooijmans, R. P., Dawson, H. D., Dehais, P., De Sapio, F., Dibbits, B., Drou, N., Du, Z. Q., Eversole, K., Fadista, J., Fairley, S., Faraut, T., Faulkner, G. J., Fowler, K. E., Fredholm, M., Fritz, E., Gilbert, J. G., Giuffra, E., Gorodkin, J., Griffin, D. K., Harrow, J. L., Hayward, A., Howe, K., Hu, Z. L., Humphray, S. J., Hunt, T., Hornshoj, H., Jeon, J. T., Jern, P., Jones, M., Jurka, J., Kanamori, H., Kapetanovic, R., Kim, J., Kim, J. H., Kim, K. W., Kim, T. H., Larson, G., Lee, K., Lee, K. T., Leggett, R., Lewin, H. A., Li, Y., Liu, W., Loveland, J. E., Lu, Y., Lunney, J. K., Ma, J., Madsen, O., Mann, K., Matthews, L., McLaren, S., Morozumi, T., Murtaugh, M. P., Narayan, J., Nguyen, D. T.,

Ni, P., Oh, S. J., Onteru, S., Panitz, F., Park, E. W., Park, H. S., Pascal, G., Paudel, Y., Perez-Enciso, M., Ramirez-Gonzalez, R., Reecy, J. M., Rodriguez-Zas, S., Rohrer, G. A., Rund, L., Sang, Y., Schachtschneider, K., Schraiber, J. G., Schwartz, J., Scobie, L., Scott, C., Searle, S., Servin, B., Southey, B. R., Sperber, G., Stadler, P., Sweedler, J. V., Tafer, H., Thomsen, B., Wali, R., Wang, J., Wang, J., White, S., Xu, X., Yerle, M., Zhang, G., Zhang, J., Zhang, J., Zhao, S., Rogers, J., Churcher, C. et Schook, L. B. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, 491(7424) :393–8. (cf. page 7).

Guan, Y. et Stephens, M. (2008). Practical Issues in Imputation-Based Association Mapping. *PLOS Genetics*, 4(12) :e1000279. (cf. page 30).

Guan, Y. et Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3) :1780–1815. (cf. page 31).

Gulko, B., Hubisz, M. J., Gronau, I. et Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics*, 47(3) :276–283. (cf. page 53).

Günther, T. et Coop, G. (2013). Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, 195(1) :205–220. (cf. page 51).

Haley, C. S. et Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity (Edinb)*, 69(4) :315–24. (cf. page 24).

Henderson, C. R., Kempthorne, O., Searle, S. R. et von Krosigk, C. M. (1959). The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, 15(2) : 192–218. (cf. page 1).

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. et Eskin, E. (2014). Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics*, 198(2) :497–508. (cf. page 48).

Hospital, F., Dillmann, C. et Melchinger, A. E. (1996). A general algorithm to compute multilocus genotype frequencies under various mating systems. *Comput Appl Biosci*, 12(6) :455–62. (cf. pages 20 and 25).

International Hapmap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164) :851–61. (cf. page 26).

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., Wu, C., Muzny, D. M., Li, Y., Zhang, W., Stanton, J. A., Brauning, R., Barris, W. C., Hourlier, T., Aken, B. L., Searle, S. M., Adelson, D. L., Bian, C., Cam, G. R., Chen, Y., Cheng, S., DeSilva, U., Dixen, K., Dong, Y., Fan, G., Franklin, I. R., Fu, S., Fuentes-Utrilla, P., Guan, R., Highland, M. A., Holder, M. E., Huang, G., Ingham, A. B., Jhangiani, S. N., Kalra, D., Kovar, C. L., Lee, S. L., Liu, W., Liu, X., Lu, C., Lv, T., Mathew, T., McWilliam, S., Menzies, M., Pan, S., Robelin, D., Servin, B., Townley, D., Wang, W., Wei, B., White, S. N., Yang, X., Ye, C., Yue, Y., Zeng, P., Zhou, Q., Hansen, J. B., Kristiansen, K., Gibbs, R. A., Flicek, P., Warkup, C. C., Jones, H. E., Oddy, V. H., Nicholas, F. W., McEwan,

J. C., Kijas, J. W., Wang, J., Worley, K. C., Archibald, A. L., Cockett, N., Xu, X., Wang, W. et Dalrymple, B. P. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science*, 344(6188) :1168–73. (cf. page 10).

Johnston, S. E., Bérénos, C., Slate, J. et Pemberton, J. M. (2016). Conserved Genetic Architecture Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep (Ovis aries). *Genetics*, 203(1) :583–598. (cf. pages 18, 33, and 54).

Kadri, N. K., Harland, C., Faux, P., Cambisano, N., Karim, L., Coppieters, W., Fritz, S., Mullaart, E., Baurain, D., Boichard, D., Spelman, R., Charlier, C., Georges, M. et Druet, T. (2016). Coding and noncoding variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B affect recombination rate in cattle. *Genome Research*, 26(10) :1323–1332. (cf. page 54).

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J. A. C., Baurain, D., Cambisano, N., Davis, S. R., Farnir, F., Grisart, B., Harris, B. L., Keehan, M. D., Littlejohn, M. D., Spelman, R. J., Georges, M. et Coppieters, W. (2011). Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nature Genetics*, 43(5) :405–413. (cf. page 47).

Kass, R. et Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90 :773–795. (cf. page 29).

Kelleher, J., Wong, Y., Albers, P. K., Wohns, A. W. et McVean, G. (2018). Inferring the ancestry of everyone. *bioRxiv*, page 458067. (cf. page 57).

Key, F. M., Abdul-Aziz, M. A., Mundry, R., Peter, B. M., Sekar, A., D'Amato, M., Dennis, M. Y., Schmidt, J. M. et Andrés, A. M. (2018). Human local adaptation of the TRPM8 cold receptor along a latitudinal cline. *PLOS Genetics*, 14(5) :e1007298. (cf. page 49).

Lander, E. S. et Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1) :185–99. (cf. page 24).

Lange, J., Yamada, S., Tischfield, S. E., Pan, J., Kim, S., Zhu, X., Socci, N. D., Jasin, M. et Keeney, S. (2016). The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. *Cell*, 167(3) :695–708.e16. (cf. page 55).

Laval, G., SanCristobal, M. et Chevalet, C. (2002). Measuring genetic distances between breeds : use of some distances in various short term evolution models. *Genetics Selection Evolution*, 34(4) :481. (cf. page 41).

Lecomte, L., Duffe, P., Buret, M., Servin, B., Hospital, F. et Causse, M. (2004). Marker-assisted introgression of five QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds. *Theor Appl Genet*, 109(3) :658–68. (cf. page 21).

Lewontin, R. C. et Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1) :175–195. (cf. page 41).

Li, N. et Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4) :2213–33. (cf. pages 1, 15, 16, 20, and 30).

Liu, X., Li, Y. I. et Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, 177(4) :1022–1034.e6. (cf. page 33).

Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., Bickhart, D. M., Cole, J. B., Null, D. J., Liu, G. E., Da, Y. et Wiggans, G. R. (2015). Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLOS Genetics*, 11(11) :e1005387. (cf. page 54).

Meuwissen, T. H. E., Hayes, B. J. et Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157(4) :1819–1829. (cf. page 1).

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(5746) :321–324. (cf. page 2).

Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, Ó., Stefánsson, K. et Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(4) :695–715. (cf. page 39).

Oget, C., Servin, B. et Palhière, I. (2019). Genetic diversity analysis of French goat populations reveals selective sweeps involved in their differentiation. *Animal Genetics*, 50(1) :54–63. (cf. page 45).

Paris, C., Servin, B. et Boitard, S. (2019). Inference of selection from genetic time series using various parametric approximations to the Wright-Fisher model. *bioRxiv*, page 696955. (cf. pages 52 and 53).

Petit, M., Astruc, J.-M., Sarry, J., Drouilhet, L., Fabre, S., Moreno, C. R. et Servin, B. (2017). Variation in recombination rate and its genetic determinism in sheep populations. *Genetics*, 207(2) :767–784. (cf. pages 13, 14, 16, 20, 32, and 54).

Petit, M., Larroque, H., Astruc, J.-M., Chassier, M., Servin, B. et Moreno, C. R. (2018a). Comparison of imputation accuracy using different snps densities and snp selection strategies based on a physical map or on a genetic map. *Proceedings of the World Congress on Genetics Applied to Livestock Production*, Electronic Poster Session - Genetic Gain - Genotyping & Phenotyping Strategies :512. (cf. page 21).

Petit, M., Larroque, H., Astruc, J.-M., Chassier, M., Servin, B. et Moreno, C. R. (2018b). Comparison of imputation accuracy using different SNPs densities and SNP selection strategies based on a physical map or on a genetic map. *In Proceedings of the World Congress on Genetics Applied to Livestock Production*, volume Genetic Gain - Genotyping & Phenotyping Strategies, page 512. (cf. page 54).

Petkova, D., Novembre, J. et Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, 48(1) :94–100. (cf. page 57).

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y. et Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48(7) :709–717. (cf. page 33).

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y. et DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10) :983–987. (cf. page 57).

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V. et Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science*, 346(6211) : 1256442. (cf. page 55).

Pritchard, J. K., Stephens, M. et Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2) :945–959. (cf. pages 1 and 56).

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P. et Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10) :1720–1728. (cf. page 57).

Reynolds, J., Weir, B. S. et Cockerham, C. C. (1983). Estimation of the coancestry coefficient : basis for a short-term genetic distance. *Genetics*, 105(3) :767–779. (cf. page 41).

Risch, N. et Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273 :1516–1517. (cf. page 25).

Rochus, C. M., Tortereau, F., Plisson-Petit, F., Restoux, G., Moreno-Romieux, C., Tosser-Klopp, G. et Servin, B. (2018). Revealing the selection history of adaptive loci using genome-wide scans for selection : An example from domestic sheep. *BMC Genomics*, 19(1) :71. (cf. pages 16, 33, 45, 46, 48, and 54).

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly,

P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R. et Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164) :913–918. (cf. page 2).

Saitou, N. et Nei, M. (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4) :406–425. (cf. page 41).

Sánchez-Villagra Marcelo R., Geiger Madeleine et Schneider Richard A. (2016). The taming of the neural crest : A developmental perspective on the origins of morphological covariation in domesticated mammals. *Royal Society Open Science*, 3(6) :160107. (cf. page 49).

Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C. et Georges, M. (2012). Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genet.*, 8(7) :e1002854. (cf. pages 14 and 54).

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. *Am J Hum Genet*, 78(4) :629–44. (cf. pages 2, 30, 42, and 43).

Sella, G. et Barton, N. H. (2019). Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annual review of genomics and human genetics*. (cf. page 57).

Servin, B. (2005). Using markers to reduce the variation in the genomic composition in marker-assisted backcrossing. *Genet Res*, 85(2) :151–7. (cf. page 21).

Servin, B., de Givry, S. et Faraut, T. (2010). Statistical confidence measures for genome maps : application to the validation of genome assemblies. *Bioinformatics*, 26(24) :3035–3042. (cf. page 5).

Servin, B., Dillmann, C., Decoux, G. et Hospital, F. (2002). MDM : a program to compute fully informative genotype frequencies in complex breeding schemes. *J Hered*, 93(3) :227–8. (cf. pages 21 and 25).

Servin, B., Faraut, T., Iannuccelli, N., Zelenika, D. et Milan, D. (2012). High-resolution autosomal radiation hybrid maps of the pig genome and their contribution to the genome sequence assembly. *BMC Genomics*, 13 :585. (cf. pages 7 and 8).

Servin, B. et Hospital, F. (2002). Optimal positioning of markers to control genetic background in marker-assisted backcrossing. *J Hered*, 93(3) :214–7. (cf. page 21).

Servin, B., Martin, O. C., Mézard, M. et Hospital, F. (2004). Toward a Theory of Marker-Assisted Gene Pyramiding. *Genetics*, 168(1) :513–523. (cf. page 21).

Servin, B. et Stephens, M. (2007). Imputation-based analysis of association studies : candidate regions and quantitative traits. *PLoS Genet*, 3(7) :e114. (cf. pages 26, 28, and 29).

Sheehan, S. et Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, 12(3) :e1004845. (cf. page 57).

Speidel, L., Forest, M., Shi, S. et Myers, S. (2019). A method for genome-wide genealogy estimation for thousands of samples. *bioRxiv*, page 550558. (cf. pages 1 and 57).

Stephens, M. (2017). False discovery rates : A new deal. *Biostatistics*, 18(2) :275–294. (cf. pages 1 and 57).

Storey, J. D. et Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16) :9440–9445. (cf. pages 1 and 46).

Sun, L., Craiu, R. V., Paterson, A. D. et Bull, S. B. (2006). Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6) :519–530. (cf. page 49).

Tataru, P., Bataillon, T. et Hobolth, A. (2015). Inference under a wright-fisher model using an accurate beta approximation. *Genetics*, 201(3) :1133–1141. (cf. page 52).

Thabuis, A., Palloix, A., Servin, B., Daubèze, A., Signoret, P., Hospital, F. et Lefebvre, V. (2004). Marker-assisted introgression of 4 phytophthora capsici resistance qtl alleles into a bell pepper line : validation of additive and epistatic effects. *Molecular Breeding*, 14(1) :9–20. (cf. page 21).

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571) :68–74. (cf. page 26).

The Bovine HapMap Consortium (2009). Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science (New York, N.Y.)*, 324(5926) :528–532. (cf. page 2).

The International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921. (cf. page 1).

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145) :661–678. (cf. page 2).

Tortereau, F., Servin, B., Frantz, L., Megens, H. J., Milan, D., Rohrer, G., Wiedmann, R., Beever, J., Archibald, A. L., Schook, L. B. et Groenen, M. A. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*, 13 :586. (cf. pages 9 and 13).

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E.,

Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. et Zhu, X. (2001). The Sequence of the Human Genome. *Science*, 291(5507) :1304–1351. (cf. page 1).

Wang, W. (2017). *Applications of Adaptive Shrinkage in Multiple Statistical Problems*. Thèse de doctorat, University of Chicago. (cf. page 57).

Wilkins, A. S., Wrangham, R. W. et Fitch, W. T. (2014). The "Domestication Syndrome" in Mammals : A Unified Explanation Based on Neural Crest Cell Behavior and Genetics. *Genetics*, 197(3) :795–808. (cf. page 49).

Xie, C., Gessler, D. D. et Xu, S. (1998). Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. *Genetics*, 149(2) : 1139–46. (cf. pages 24 and 25).

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. et Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7) :565–569. (cf. page 2).

Zhong, S. et Jannink, J.-L. (2007). Using Quantitative Trait Loci Results to Discriminate Among Crosses on the Basis of Their Progeny Mean and Variance. *Genetics*, 177(1) :567–576. (cf. page 54).

Zhou, X., Carbonetto, P. et Stephens, M. (2013). Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLOS Genetics*, 9(2) :e1003264. (cf. pages 32 and 33).

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C. P., Sonstegard, T. S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J. A. et Salzberg, S. L. (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology*, 10(4) :R42. (cf. page 2).

Curriculum Vitae

BERTRAND SERVIN

Chercheur en génétique statistique / science des données

@ bertrand.servin@inra.fr +33-5-6128-5117
INRA Genphyse, Chemin de Borde-Rouge Auzeville, 31326 Castanet-Tolosan Toulouse, France
http://genphyse.toulouse.inra.fr/people/servin/bertrand @BertrandServin
in linkedin.com/in/bertrand-servin github.com/BertrandServin orcid.org/0000-0001-5141-0913

EXPERIENCE

Chargé de Recherche

Institut National de la Recherche Agronomique - Génétique Animale

Sep 2006 – En cours Toulouse, France

Animateur de l'équipe *Dynamique des génomes et des populations* au laboratoire GenPhySE (depuis Mars 2016).

- Développement de méthodes statistiques pour l'analyse de données génétiques.
- Analyses de jeux de données de grande taille en génomique des populations

Post-doctorat

University of Washington - Department of Statistics

Fév 2004 – Août 2006 Seattle, WA, USA

Développement de méthodes statistiques pour les analyses d'association exploitant l'imputation de génotypes. Encadrement : Matthew Stephens

Doctorat

Institut National de la Recherche Agronomique - Biologie et Amélioration des Plantes

Déc 2000 - Déc 2003 Gif-sur-Yvette, France

Optimisation du rétrocroisement assiste par marqueurs et mise au point de plans d'expérience pour le pyramidage de gènes. Encadrement: Frédéric Hospital

RECHERCHE

Génétique de l'adaptation

- Méthodes pour la détection de l'adaptation locale.
- Études de la diversité et l'adaptation des populations d'animaux d'élevage à partir de données pan-génomiques.

Cartographie des génomes

- Intégration de données pour l'amélioration des assemblages de génomes
- Caractérisation du processus de recombinaison et de sa variabilité.

Génétique des caractères complexes

- Méthodes statistiques pour les études d'association pan-génomiques
- Déterminisme génétique de la recombinaison
- Génétique de la résistance de l'abeille domestique à *Varroa destructor*

PRODUCTION

47 publications

h-index: 23 **i10-index:** 37

COMPÉTENCES

Génétique des populations
Statistiques
Génétique quantitative
Recherche reproductible

python
numpy/scipy
R
Linux
C

LANGUES

Français ●●●●●

Anglais ●●●●○

DIPLÔMES

Habilitation à Diriger les Recherces

Université de Toulouse

Soutenance prévue le 12 Novembre 2019

Contributions à l'analyse statistique de données génétiques

Thèse de Génétique

Université Paris-Sud

Dec 2000 – Dec 2003

Méthodes de construction de génotypes assistée par marqueurs

Diplôme d'ingénieur agronome

AgroParisTech

Sept 1997 – Sep 2000

DEA Biologie, Diversité et Adaptation des Plantes Cultivées. Option: Ressources génétiques et amélioration des plantes.

FORMATION ET ENSEIGNEMENT

Enseignement

Le modèle linéaire mixte (20h)

Université Paul Sabatier - Master SID

2008-2011 Toulouse, France

Principes généraux, techniques d'estimation (ML, REML ...), tests d'hypothèses et critères d'ajustement, modèles hiérarchiques Bayésiens.

Module de Génétique Statistique (10h)

Université Paul Sabatier - Master Bioinformatique

2011-2016 Toulouse, France

Introduction du module, Déséquilibre de liaison, principes de la coalescence

Formations

Software and Statistical Methos for Population Genetics

Université Grenoble Alpes

2015,2017 Aussois, France

Formation à l'analyse de données de génétique des populations auprès de doctorants et jeunes chercheurs (1 semaine).

Analyse de données post-génomiques

Plateforme Genotoul - biostats

2012,2013,2014,2015,2017 Toulouse, France

Formation au modèle linéaire (1 journée) et modèle linéaire mixte (1 journée) auprès de doctorants et chercheurs. Sessions pratiques sur le logiciel R.

LOGICIELS

`metamap`

forge-dga.jouy.inra.fr/projects/metamap

Production de cartes robustes à partir des sorties du logiciel `carthagene`.

`hapflk`

forge-dga.jouy.inra.fr/projects/hapflk

Détection de signatures de sélection basée sur la différentiation de populations : approches simples marqueurs et haplotypiques.

JEUX DE DONNÉES

Génotypes de la population ovine Lacaune

doi:10.5281/zenodo.804264

Génotypes de 6229 animaux de la race ovine Lacaune (50K SNP).

Génotypes haute-densité des populations ovines françaises

doi:10.5281/zenodo.237116

Génotypes de 27 populations ovines françaises (600K SNP).

ENCADREMENT

Doctorants

Alice Danguy des Déserts (50%)

INRA, Clermont-Ferrand

2019 – en cours

Patrons de recombinaison chez le Blé tendre et exploitation en sélection

Klavdija Poklukar (25%)

KIS, Slovénie

2018 – en cours

Déterminismes du métabolisme des acides gras chez le porc

Cyriel Paris (50%)

INRA, Toulouse

2016 – en cours

Estimation des paramètres de sélection à partir de données génomiques temporelles

Morgane Petit (50%)

INRA, Toulouse

2014 – 2017

Génétique de la recombinaison chez le mouton domestique

Christina Rochus (25%)

SLU, Suède

2013 – 2017

Diversité génétique du mouton domestique

Jason Lapeyronnie (50%)

INRA, Toulouse

2014 – 2017

Méthodes statistiques pour la détection de l'adaptation locale

Maria-Ines Fariello (33%)

INRA, Toulouse

2009 – 2013

Méthodes statistiques pour la détection de signatures de sélection

Postdoctorants

Sonia Eynard

INRA, Toulouse

2017 – en cours

Génétique de la résistance de l'abeille domestique à *Varroa destructor*.

Productions scientifiques

Articles scientifiques

Alberto, F. J., F. Boyer, P. Orozco-terWengel, I. Streeter, B. Servin, P. de Villemereuil, B. Benjelloun et al. 2018. "Convergent Genomic Signatures of Domestication in Sheep and Goats." *Nat Commun* 9 : 813.

Amadou, C., G. Pascal, S. Mangenot, M. Glew, C. Bontemps, D. Capela, S. Carrère et al. 2008. "Genome Sequence of the Beta-Rhizobium Cupriavidus Taiwanensis and Comparative Genomics of Rhizobia." *Genome research* 18 (9) : 1472-83.

Bertolini, F., B. Servin, A. Talenti, E. Rochat, E. S. Kim, C. Oget, I. Palhière et al. 2018. "Signatures of Selection and Environmental Adaptation across the Goat Genome Post-Domestication." *Genet Sel Evol* 50 : 57.

Boitard, S., M. Boussaha, A. Capitan, D. Rocha et B. Servin. 2016a. "Uncovering Adaptation from Sequence Data : Lessons from Genome Resequencing of Four Cattle Breeds." *Genetics* 203 : 433-50.

Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott et M. SanCristobal. 2010b. "Detecting Selection in Population Trees : The Lewontin and Krakauer Test Extended". *Genetics* 186 (1) : 241-262.

Bonnet, A., B. Servin, P. Mulsant et B. Mandon-Pepin. 2015. "Spatio-Temporal Gene Expression Profiling during In Vivo Early Ovarian Folliculogenesis : Integrated Transcriptomic Study and Molecular Signature of Early Follicular Growth." *PLoS One* 10 : e0141482.

Bouchet, S., B. Servin, P. Bertin, D. Madur, V. Combes, F. Dumas, D. Brunel, J. Laborde, A. Charcosset et S. Nicolas. 2013. "Adaptation of Maize to Temperate Climates : Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus". *Plos One* 8 (8). doi :10 . 1371 / journal.pone.0071377.

Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana et al. 2018. "Meta-Analysis of Genome-Wide Association Studies for Cattle Stature Identifies Common Genes That Regulate Body Size in Mammals." *Nat Genet* 50 : 362-367.

Colli, L., M. Milanesi, A. Talenti, F. Bertolini, M. Chen, A. Crisà, K. G. Daly et al. 2018. "Genome-Wide SNP Profiling of Worldwide Goat Populations Reveals Strong Partitioning of Diversity and Highlights Post-Domestication Migration Routes." *Genet Sel Evol* 50 : 58.

Crepieux, S., C. Lebreton, B. Servin et G. Charmet. 2004a. "IBD-Based QTL Detection in Inbred Pedigrees : A Case Study of Cereal Breeding Programs". *Euphytica* 137 (1) : 101.

Crepieux, S., C. Lebreton, B. Servin et G. Charmet. 2004b. "Quantitative Trait Loci (QTL) Detection in Multicross Inbred Designs : Recovering QTL Identical-by-Descent Status Information From Marker Data". *Genetics* 168 (3) : 1737-1749.

Darrier, B., H. Rimbert, F. Balfourier, L. Pingault, A. Josselin, B. Servin, J. Navarro, F. Choulet, E. Paux et P. Sourdille. 2017. "High-Resolution Mapping of Crossover Events in the Hexaploid Wheat Genome Suggests a Universal Recombination Mechanism". *Genetics* 206 : 1373-1388.

Demars, J., M. Cano, L. Drouilhet, F. Plisson-Petit, P. Bardou, S. Fabre, B. Servin et al. 2017. "Genome-Wide Identification of the Mutation Underlying Fleece Variation and Discriminating Ancestral Hairy Species from Modern Woolly Sheep." *Mol Biol Evol* 34 : 1722-1729.

Dong, Y., M. Xie, Y. Jiang, N. Xiao, X. Du, W. Zhang, G. Tosser-Klopp et al. 2013. "Sequencing and Automated Whole-Genome Optical Mapping of the Genome of a Domestic Goat (Capra Hircus)". *Nature Biotechnology* 31 : 135-41.

Du, X., B. Servin, J. Womack, C. Cao, M. Yu, Y. Dong, W. Wang et S. Zhao. 2014. "An Update of the Goat Genome Assembly Using Dense Radiation Hybrid Maps Allows Detailed Analysis of Evolutionary Rearrangements in Bovidae." *BMC Genomics* 15 : 625.

Faraut, T., S. de Givry, C. Hitte, Y. Lahbib-Mansais, M. Morisson, D. Milan, T. Schiex, B. Servin, A. Vignal, F. Galibert et M. Yerle. 2009. "Contribution of Radiation Hybrids to Genome Mapping in Domestic Animals." *Cytogenet Genome Res* 126 : 21-33.

Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal et B. Servin. 2013. "Detecting Signatures of Selection through Haplotype Differentiation among Hierarchically Structured Population". *Genetics* 193 : 929-941.

Fariello, M. I., B. Servin, G. Tosser-Klopp, R. Rupp, C. R. Moreno, I. S. G. Consortium, M. San Cristobal et S. Boitard. 2014. "Selection Signatures in Worldwide Sheep Populations". *PLoS One* 9 (8) : 1-12.

Frésard, L., S. Leroux, P. Dehais, B. Servin, H. Gilbert, O. Bouchez, C. Klopp et al. 2012. "Fine Mapping of Complex Traits in Non-Model Species : Using next Generation Sequencing and Advanced Intercross Lines in Japanese Quail." *BMC Genomics* 13 (october) : 551.

Frésard, L., S. Leroux, B. Servin, D. Gourichon, P. Dehais, M. S. Cristobal, N. Marsaud et al. 2014. "Transcriptome-Wide Investigation of Genomic Imprinting in Chicken." *Nucleic Acids Res* 42 : 3768-82.

Gholami, M., C. Reimer, M. Erbe, R. Preisinger, A. Weigend, S. Weigend, B. Servin et H. Simianer. 2015. "Genome Scan for Selection in Structured Layer Chicken Populations Exploiting Linkage Disequilibrium Information". *PLoS One* 10 (7) : 1-15.

Groenen, M. A. M., A. L. Archibald, H. Uenishi, C. K. Tuggle, Y. Takeuchi, M. F. Rothschild, C. Rogel-Gaillard et al. 2012. "Analyses of Pig Genomes Provide Insight into Porcine Demography and Evolution". *Nature* 491 : 393-398.

Hamzic, E., B. Bed, H. Juin, R. Hawken, M. S. Abrahamsen, J. M. Elsen, B. Servin, M. H. Pinard-van der Laan et O. Demeure. 2015. "Large-Scale Investigation of the Parameters in Response to Challenge in Broilers." *J Anim Sci* 93 : 1830-40.

Hamzić, E., B. Buitenhuis, F. Hérault, R. Hawken, M. S. Abrahamsen, B. Servin, J.-M. Elsen, M.-H. Pinard-van der Laan et B. Bed. 2015. "Genome-Wide Association Study and Biological Pathway Analysis of the Eimeria Maxima Response in Broilers." *Genet Sel Evol* 47 : 91.

Jiang, Y., M. Xie, W. Chen, R. Talbot, J. F. Maddox, T. Faraut, C. Wu et al. 2014. "The Sheep Genome Illuminates Biology of the Rumen and Lipid Metabolism." *Science* 344 : 1168-73.

Kijas, J. W., J. A. Lenstra, B. Hayes, S. Boitard, L. R. Porto Neto, M. S. Cristobal, B. Servin et al. 2012. "Genome-Wide Analysis of the Worlds Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection". *PLoS Biol* 10 (2) : e1001258.

Lecomte, L., P. Duffé, M. Buret, B. Servin, F. Hospital et M. Causse. 2004. "Marker-Assisted Introgression of Five QTLs Controlling Fruit Quality Traits into Three Tomato Lines Revealed Interactions between QTLs and Genetic Backgrounds". *Theoretical and Applied Genetics* 109 (3) : 658-668.

Marty, A., Y. Amigues, B. Servin, G. Renand, H. Levéziel et D. Rocha. 2010. "Genetic Variability and Linkage Disequilibrium Patterns in the Bovine DNAJA1 Gene." *Molecular biotechnology* 44 (3) : 190-7.

Monestier, O., B. Servin, S. Auclair, T. Bourquard, A. Poupon, G. Pascal et S. Fabre. 2014. "Evolutionary Origin of Bone Morphogenetic Protein 15 and Growth and Differentiation Factor 9 and Differential Selective Pressure between Mono- and Polyovulating Species." *Biol Reprod* 91 : 83.

Oget, C., B. Servin et I. Palhière. 2019. "Genetic Diversity Analysis of French Goat Populations Reveals Selective Sweeps Involved in Their Differentiation". *Animal Genetics* 50 (1) : 54-63.

Paris, C., B. Servin et S. Boitard. 2019. "Inference of Selection from Genetic Time Series Using Various Parametric Approximations to the Wright-Fisher Model". *bioRxiv :* 696955.

Petit, M., J.-M. Astruc, J. Sarry, L. Drouilhet, S. Fabre, C. R. Moreno et B. Servin. 2017a. "Variation in Recombination Rate and Its Genetic Determinism in Sheep Populations." *Genetics* 207 : 767-784.

Riquet, J., H. Gilbert, B. Servin, M.-P. Sanchez, N. Iannuccelli, Y. Billon, J.-P. Bidanel et D. Milan. 2011. "A Locally Congenic Backcross Design in Pig : A New Regional Fine QTL Mapping Approach Miming Congenic Strains Used in Mouse." *BMC Genet* 12 (january) : 6.

Rochus, C. M., F. Tortereau, F. Plisson-Petit, G. Restoux, C. Moreno-Romieux, G. Tosser-Klopp et B. Servin. 2018. "Revealing the Selection History of Adaptive Loci Using Genome-Wide Scans for Selection : An Example from Domestic Sheep". *BMC Genomics* 19 (1) : 71.

Rousseau, S., N. Iannuccelli, M.-J. Mercat, C. Naylies, J.-C. Thouly, B. Servin, D. Milan, E. Pailhoux et J. Riquet. 2013. "A Genome-Wide Association Study Points out the Causal Implication of SOX9 in the Sex-Reversal Phenotype in XX Pigs". *PLoS One* 8 (11) : e79882.

Roux, P.-F., S. Boitard, Y. Blum, B. Parks, A. Montagner, E. Mouisel, A. Djari et al. 2015. "Combined QTL and Selective Sweep Mappings with Coding SNP Annotation and Cis-eQTL Analysis Revealed PARK2 and JAG2 as New Candidate Genes for Adiposity Regulation." *G3 (Bethesda)* 5 : 517-29.

Sanchez, M.-P., T. Tribout, N. Iannuccelli, M. Bouffaud, B. Servin, A. Tenghe, P. Dehais, N. Muller, M. P. Del Schneider, M.-J. Mercat, C. Rogel-Gaillard, D. Milan, J.-P. Bidanel et H. Gilbert. 2014. "A Genome-Wide Association Study of Production Traits in a Commercial Population of Large White Pigs : Evidence of Haplotypes Affecting Meat Quality." *Genetics, selection, evolution : GSE* 46 (1) : 12.

Servin, B., C. Dillmann, G. Decoux et F. Hospital. 2002. "MDM : A Program to Compute Fully Informative Genotype Frequencies in Complex Breeding Schemes". *Journal of Heredity* 93 (3) : 227-228.

Servin, B. 2005b. "Using Markers to Reduce the Variation in the Genomic Composition in Marker-Assisted Backcrossing." *Genetical Research* 85 (2) : 151-157.

Servin, B., S. de Givry et T. Faraut. 2010b. "Statistical Confidence Measures for Genome Maps : Application to the Validation of Genome Assemblies". *Bioinformatics* 26 (24) : 3035-3042.

Servin, B., T. Faraut, N. Iannuccelli, D. Zelenika et D. Milan. 2012. "High-Resolution Autosomal Radiation Hybrid Maps of the Pig Genome and Their Contribution to the Genome Sequence Assembly." *BMC Genomics* 13 : 585.

Servin, B., et F. Hospital. 2002. "Optimal Positioning of Markers to Control Genetic Background in Marker-Assisted Backcrossing." *Journal of Heredity* 93 (3) : 214-217.

Servin, B., O. Martin, M. Mézard et F. Hospital. 2004. "Toward a Theory of Marker-Assisted Gene Pyramiding." *Genetics* 168 (1) : 513-523.

Servin, B., et M. Stephens. 2007. "Imputation-Based Analysis of Association Studies : Candidate Regions and Quantitative Traits". *PLOS Genetics* 3 (7) : e114.

Thabuis, A., A. Palloix, B. Servin, A. Daubèze, P. Signoret, F. Hospital et V. Lefebvre. 2004. "Marker-Assisted Introgression of 4 Phytophthora Capsici Resistance QTL Alleles into a Bell Pepper Line : Validation of Additive and Epistatic Effects". *Molecular Breeding* 14 (1) : 9-20.

Tortereau, F., C. R. Moreno, G. Tosser-Klopp, B. Servin et J. Raoul. 2017. "Development of a SNP Panel Dedicated to Parentage Assignment in French Sheep Populations." *BMC Genet* 18 : 50.

Tortereau, F., B. Servin, L. Frantz, H.-J. Megens, D. Milan, G. Rohrer, R. Wiedmann, J. Beever, A. L. Archibald, L. Schook et M. Groenen. 2012. "A High Density Recombination Map of the Pig Reveals a Correlation between Sex-Specific Recombination and GC Content". *BMC Genomics* 13 (1) : 586.

Actes de colloques

Boitard, S. 2010. "SNP Selection Using Sparse PLS". In *Proceedings of the World Congress on Genetics Applied to Livestock Production,* avec la collaboration de B. Sansas, B. Servin et M. SanCristobal, t. Methods and tools : Population genomics, 0160. Leipzig, Germany.

Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. M. Abdallah, S. Blott et M. San Cristobal. 2010a. "Detecting Selection in Population Trees : An Extension of the Lewontin and Krakauer Test with an Application to Pig." In *Proceedings of the World Congress on Genetics Applied to Livestock Production,* t. Methods and tools : Population genomics, 0113.

Hamzic, E., B. Bed'Hom, H. Juin, R. Hawken, M. S. Abrahamsen, J.-M. Elsen, B. Servin, M.-H. P.-v. der Laan et O. Demeure. 2014. "Plasma Components as Traits for Resistance to Coccidiosis in Chicken". In *Proceedings of the World Congress on Genetics Applied to Livestock Production,* t. Genetics of Trait Complexes : Disease resistance, 098. Vancouver, Canada.

Oget, C., P. Martin, B. Servin et I. Palhière. 2018. "Signatures Génomiques de l'histoire Évolutive Des Races Caprines Françaises". In *24e Rencontres Recherche Ruminants,* 106-109. Paris, France.

Petit, M., S. Fabre, J. Sarry, C. Moreno-Romieux et B. Servin. 2017b. "Insights on the Genetic Determinism and Evolution of Recombination Rates by Combining Multiple Genome-Wide Datasets in the Sheep". Published : 68. Annual Meeting of the European Associationfor Animal Production (EAAP). Tallin, Estonia. `https://hal.archives-ouvertes.fr/hal-01604584`.

Petit, M., H. Larroque, J.-M. Astruc, M. Chassier, B. Servin et C. R. Moreno. 2018. "Comparison of Imputation Accuracy Using Different SNPs Densities and SNP Selection Strategies Based on a Physical Map or on a Genetic Map". In *Proceedings of the World Congress on Genetics Applied to Livestock Production,* t. Genetic Gain - Genotyping & Phenotyping Strategies, 512. Auckland, New-Zealand.

Saintilan, R., M.-N. Fouilloux, A. Capitan, B. Servin, T. Tribout, E. Venot et F. Phocas. 2014. "Genetic Architecture of Birth and WeaningTraits In Charolais Beef Cattle". In *Proceedings of the World Congress on Genetics Applied to Livestock Production,* t. Species Breeding : Beef cattle, 738. Vancouver, Canada.

Servin, B., S. de Givry et T. Faraut. 2010a. “Modelling Radiation Hybrid Maps Uncertainty and Application to the Validation of Whole Genome Assemblies”. In *Proceedings of the World Congress on Genetics Applied to Livestock Production,* t. Methods and tools : Software and bioinformatics, 0245. Leipzig, Germany.

Servin, B., S. Boitard, C. Chevalet, M.-I. Fariello, F. Phocas et M. SanCristobal. 2014a. “Accounting for Population Structure and Haplotype Diversity in Whole Genome Scans for Selection Signatures”. In *Proceedings of the World Congress on Genetics Applied to Livestock Production,* t. Methods and Tools : Statistical and genomic tools for mapping QTL and genes, 674. Vancouver, Canada.

Servin, B., O. Martin, M. Mézard et F. Hospital. 2003. “A General Algorithm for the Optimization of Marker Assisted Gene Pyramiding”. Coruña, Spain : Conselleria de Politica Agroalimentaria e Desenvolvemento Rural. Xunta de Galicia.

Servin, B., C. Rogel-Gaillard, D. Milan, J.-P. Bidanel et J. Riquet. 2014b. “Le séquençage du génome du porc : apport des nouvelles technolgies de séquençage à la génomique et la génétique porcine.” In *46e Journées de la Recherche Porcine,* 7-18. Paris, France : IFIP.

Présentations orales, posters et séminaires

Boitard, S. 2015. “Disentangling Demography and Selection Effects of Cattle Domestication : New Insights from 1000 Bull Genomes Project”. Avec la collaboration de M. Dolezal, B. Servin, D. Fischer, J. Decker, I. Mac Leod, Q. Zhang, B. Guldbrandtsen, M. S. Lund, A. Bagnato et J. Vikki. Annual Meeting of the Society for Molecular Biology and Evolution, Vienna, Austria.

———. 2019. “Detecting Selection from Genomic Time Series : The Beta with Spikes Approximation”. Avec la collaboration de C. Paris et B. Servin. Annual Meeting of the Society for Molecular Biology and Evolution, Manchester, UK.

Demars, J. 2017. “EIF2S2 Retroposition into IRF2BP2 Underlies Fleece Variation in Sheep through the Formation of a Long Endogenous Double-Stranded RNA”. Avec la collaboration de M. Cano, L. Drouilhet, S. Fabre, B. Servin, P. Mulsant, G. Tosser-Klopp et D. Allain. XXV International Plant and Animal Genome, San Diego, CA, USA.

Eynard, S. 2018. “Explaining the Resistance to Varroa Destructor in the French Honey Bee Population with VSH, SMR and Colony Dynamic Data”. Avec la collaboration de B. Basso, A.-L. Guirao, F. Mondet, A. Vignal et B. Servin. European Conference of Apidology, Ghent, Belgium.

———. 2019. “Characterisation of Honey Bee Colonies Genotypes from Pool Sequences”. Avec la collaboration de C. Sann, F. Mondet, B. Basso, K. Canale-Tabet, R. Mahla, O. Bouchez, Y. Leconte, Y. Poquet, F. Phocas, F. Guillaume, A. Decourtye, L. Genestout, A. Vignal et B. Servin. ApiMondia, Montreal, QC, Canada.

Frésard, L. 2014. “Trasncriptome Wide Investigation of Genomic Imprinting in Chicken”. Avec la collaboration de S. Leroux, B. Servin, D. Gourichon, P. Dehais, M. San Cristobal, N. Marsaud, F. Vignoles et al. San Diego, CA, USA.

Hamzic, E. 2014. “Genome-Wide Association Study for Response to Elmeria Maxima Challegne in Broilers”. Avec la collaboration de B. Bed’Hom, F. Herault, H. Juin, R. Hawken, M. S. Abrahamsen, J.-M. Elsen, B. Servin et M. H. Pinard-van der Laan. XXII Internatinal Plant and Animal Genome, San Diego, CA, USA.

Paris, C. 2019. “Comparison of Different Models to Infer Selection from Genomic Time Series Data”. Avec la collaboration de B. Servin et S. Boitard. Mathematical and Computational Evolutionary Biology, Porquerolles, France.

Petit, M. 2016. “Fine-Scale Recombination Maps in the Sheep”. Avec la collaboration de C. R. Moreno et B. Servin. XXIV International Plant and Animal Genome, San Diego, CA, USA.

Rochus, C. M. 2016. “High Density Genome Sca for Selection in French Sheep”. Avec la collaboration de F. Tortereau, C. R. Moreno, G. Tosser-Klopp et B. Servin. XXIV Internation Plant and Animal Genome, San Diego, CA, USA.

Servin, B. 2001. “Precision Graphical Genotypes in Backcrossing”. Avec la collaboration de F. Hospital. 7th Quantitative Trait Locus Mapping and Marker-Assisted Selection Workshop, Valencia, Spain.

———. 2005a. “A Bayesian Approach to the Analysis of Candidate Genes Association Studies”. Avec la collaboration de M. Stephens. American Society of Human Genetics, Salt Lake City, UT, USA.

———. 2006. “Efficient Multipoint Analysis of Association Studies.” Avec la collaboration de P. Scheet et M. Stephens. American Society of Human Genetics, New Orleans, LA, USA.

———. 2012a. “Utilisation d’un Modèle Du Déséquilibre de Liaison Pour La Détection de Traces de Sélection Dans Des Populations Structurées”. Avec la collaboration de M.-I. Fariello, S. Boitard et M. San Cristobal. Séminaire Du Laboratoire CBGP, Montpellier, France.

———. 2012b. “Utilisation d’un Modèle Du Déséquilibre de Liaison Pour La Détection de Traces de Sélection Dans Des Populations Structurées”. Avec la collaboration de M.-I. Fariello, S. Boitard et M. San Cristobal. Séminaire Du Laboratoire EDB, Toulouse, France.

———. 2013. “A New Method for the Detection of Selection Signatures”. Avec la collaboration de M. I. Fariello, S. Boitard, C. Chevalet, M. San Cristobal et M. Bonhomme. Synbreed Project Meeting, Göttingen, Germany.

———. 2014a. “Detection Selection Signatures from Population Differentiation : Results from the Sheephapmap and 1000 Bull Genomes Datasets”. Avec la collaboration de M. San Cristobal, A. Capitan, D. Rocha et S. Boitard. XXII Internation Plant and Animal Genome, San Diego, CA, USA.

Servin, B. 2014b. “On the Advantages of a Multipoint Approach for the Detection of Selection Signatures”. Avec la collaboration de S. Boitard. Mathematical and Computational Evolutionary Biology Meeting, Saint-Martin-de-Londres, France.

———. 2015. “A Multipoint Approach for the Detection of Selection Signatures”. Avec la collaboration de M. I. Fariello, C. Chevalet, M. San Cristobal et M. Bonhomme. SMBE Sattelite Meeting : Biology of Adaptation, Saint-Martin-de-Londres, France.

———. 2016a. “Selection Signatures in the Cattle Genome : Lessons from Large Scale Resequencing of Four International Breeds”. Avec la collaboration de S. Boitard. XXIV Internation Plant and Animal Genome, San Diego, CA, USA.

———. 2016b. “Statistical Methods for the Functional Characterization of Selection Signatures”. Avec la collaboration de J. Lapeyronnie et C. Chevalet. International Conference on Quantitative Genetics, Madison, WI, USA.

———. 2017a. “Models for Genetic Differentiation of Populations and Detection of Adaptive Loci”. Séminaire de Statistiques AgroParisTech, Paris, France.

———. 2017b. “QTL Detection Using Genome Scans for Selection on Resequencing Data : Illustration in Cattle”. XXV International Plant and Animal Genome, San Diego, CA, USA.

———. 2017c. “Récit Génétique de l’histoire de La Domestication Du Mouton”. Rencontres Du Musée de l’homme, Paris, France.

———. 2018a. “Effect of Fitness Landscape, Population Structure and Linkage Disequilibirum on the Detection of Local Adaptation”. Avec la collaboration de J. Lapeyronnie et C. Chevalet. Evolution, Montpellier, France.

———. 2018b. “Impact of Population Structure, Linkage Disequilibirum and Fitness Landscape on the Dtection of Local Adaptation”. Séminaire Du Laboratoire LIPM, Toulouse, France.

———. 2019a. “Combined Approaches to Study the Impact of Eveolutoinary Pressures on Recombination in Sheep”. Séminaire Du Laboratoire GIGA, Liège, Belgique.

———. 2019b. “Détection de Locus Adaptatifs En Populations Structurées”. Séminaire de La Ferme Du Moulon, Gif-sur-Yvette, France.

Servin, B., et M.-J. Mercat. 2019. “TREASURE : Diversité Génétique et Phénotypique Des Races Locales de Porcs Européens”. Avec la collaboration de J. Riquet, K. Poklukar et C. Mestre. Journée de Restitution Du Projet TREASURE à La Filière Noir-de-Bigorre, Tarbes, France.

Vignal, A. 2018. “De Novo Genome Assembly of a Western European Apis Mellifera Mellifera Black Bee”. Avec la collaboration de S. Eynard, C. Klopp, K. Canale-Tabet, W. Marande, A. Roulet, C. Donnadieu et B. Servin. European Conference of Apidology, Ghent, Belgium. ht tps://prodinra.inra.fr/record/466650.

Publications des doctorants encadrés

Maria-Ines Fariello

— Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal et B. Servin. 2013. "Detecting Signatures of Selection through Haplotype Differentiation among Hierarchically Structured Population". Genetics 193 : 929-941.

— Fariello, M. I., B. Servin, G. Tosser-Klopp, R. Rupp, C. R. Moreno, I. S. G. Consortium, M. San Cristobal et S. Boitard. 2014. "Selection Signatures in Worldwide Sheep Populations". PLoS One 9 (8) : 1-12.

Christina Rochus

— Rochus, C. M., F. Tortereau, F. Plisson-Petit, G. Restoux, C. Moreno-Romieux, G. Tosser-Klopp et B. Servin. 2018. "Revealing the Selection History of Adaptive Loci Using Genome-Wide Scans for Selection : An Example from Domestic Sheep". BMC Genomics 19 (1) : 71.

Morgane Petit

— Petit, M., J.-M. Astruc, J. Sarry, L. Drouilhet, S. Fabre, C. R. Moreno et B. Servin. 2017a. "Variation in Recombination Rate and Its Genetic Determinism in Sheep Populations." Genetics 207 : 767-784.

Cyriel Paris

— Paris, C., B. Servin et S. Boitard. 2019. "Inference of Selection from Genetic Time Series Using Various Parametric Approximations to the Wright-Fisher Model". bioRxiv : 696955. *En Révision au journal Genes | Genomes | Genetics (G3)*

Projets financés sur Appel d'offre

Acronyme	Objectif	Source	Année	Rôle	Publications associées
DeLiSus	Diversité génétique des races porcines françaises	ANR	2007-10	Contributeur	Bonhomme et al. 2010 ; Servin et al. 2012 ; Groenen et al. 2012 ; Tortereau et al. 2012
Swan	Déterminisme des anomalies congénitales chez le porc	ANR	2008-12	Contributeur	Rousseau et al. 2013
Annosel	Annotations de signatures de sélection	INRA SelGen	2013-16	**Coordinateur**	Boitard et al. 2016 ; Bouwman et al. 2018
BoDeLiRe	Génétique de la recombinaison chez le Blé et les Ovins	INRA Selgen	2014-17	**Coordinateur**	Petit et al. 2017
Treasure	Diversité des races locales de porc et de leurs produits	H2020	2015-19	Contributeur	
VarGoats	Adaptation et domestication des chèvres domestiques	France Génomique	2015-19	Contributeur	Colli et al. 2018 ; Bertolini et al. 2018
IMAGE	Gestion innovante des ressources génétiques animales	H2020	2016-20	Contributeur	Paris et al. 2019
Beestrong	Génétique de la résistance de l'abeille à Varroa	P2SA	2016-20	Contributeur	
Haptitude	Utilisation d'haplotypes pour les études de génétique	INRA Selgen	2017-20	**Coordinateur**	Paris et al. 2019
Path2Bos	Domestication des bovins par une approche paléogénomique	ANR	2018-21	Contributeur	
Smarter	Efficacité et robustesse des petits ruminants	H2020	2018-22	**Responsable de WP**	