L'usage des points chauds de recombinaison chez le mouton et son déterminisme génétique

Héléna Vassilieff

Master 2

Bio-informatique

Parcours
Analyse et Modélisation des Données

RAPPORT DE STAGE PRESENTE PAR:
VASSILIEFF Héléna

SUJET :
L'usage des points chauds de recombinaison chez le mouton et son déterminisme génétique

Responsable du Stage:
SERVIN Bertrand

INRAE UMR genphyse
35 chemin de bordes rouge
31320 Auzeville Tolosane

08/2020

Abréviations

FDR : False discovery rate

GWAS : Genome Wide Association Study

PRDM9 : PR domain zinc finger protein 9

SNP : single nucleotide polymorphism

TN : Taux estimé Naïf

TML : Taux estimé par Maximum de Vraisemblance

Table des matières

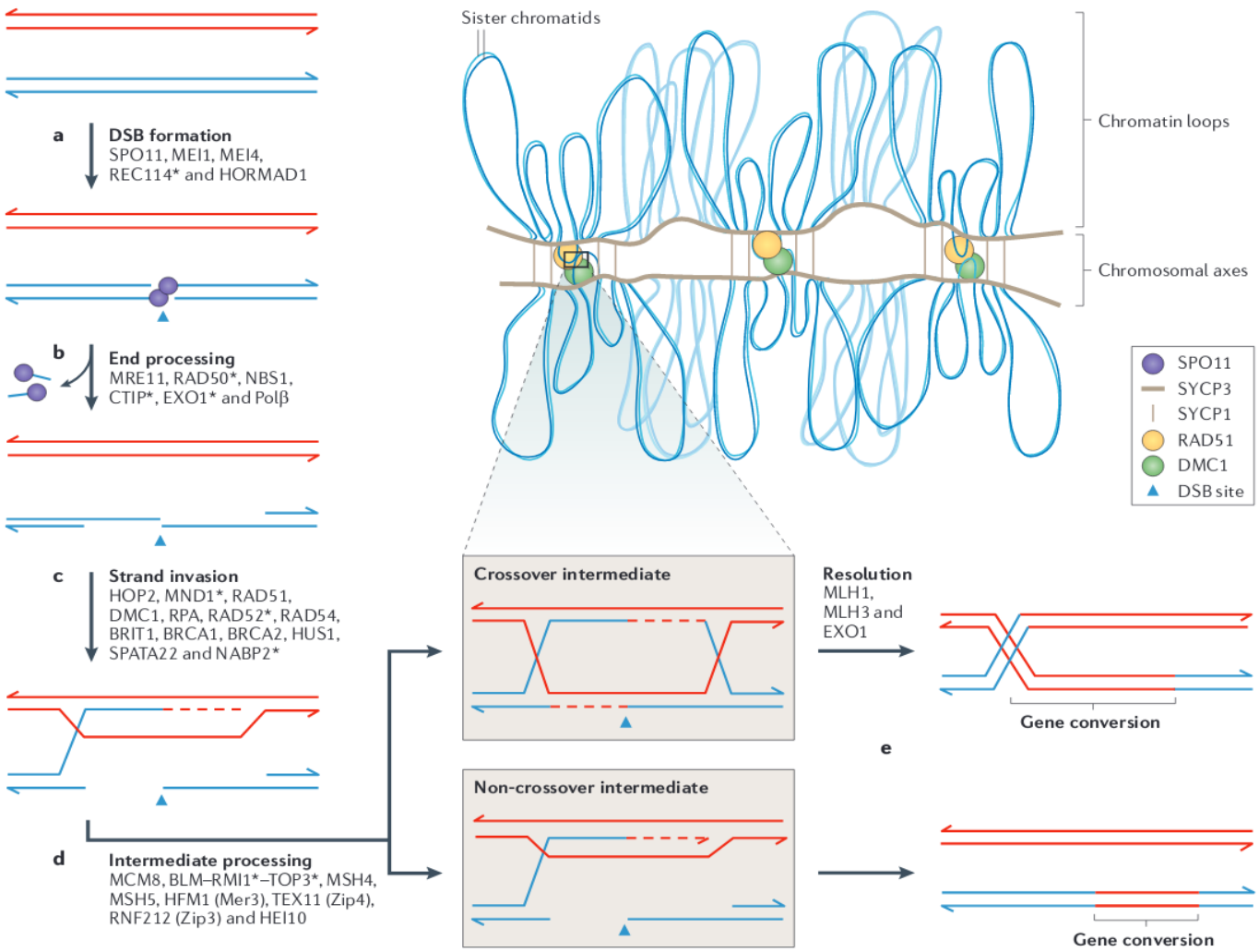


Figure 1 : processus de la recombinaison *(figure provenant de Baudat et al. 2013)*

a: des protéines comme SPO11 réalisent une cassure double brin sur des chromosomes liés par l'axe chromosomal.

b: les protéines de la cassure double brin sont relâchées et des endonucléases coupe encore du matériel génétique.

c: invasion des brins.

d: résolution des coupures par des jonctions de holliday.

e: apparition d'un crossing over ou d'un non crossing over.

* : protéines probablement impliquées dans le processus de recombinaison mais sans certitudes.

1) BIBLIOGRAPHIE

La recombinaison chez les mammifères représentent une littérature très dense. Ainsi, pour une partie des phénomènes aujourd'hui connus ce sont des revues qui serviront ici de références. Baudat et al. (2013) en ce qui concerne les mécanismes de recombinaisons chez les mammifères et leurs évolutions. Coop and Przeworski (2007) traitent plus spécifiquement de l'évolution de la recombinaison chez l'humain. Peñalba and Wolf (2020) et dans une moindre mesure Baudat et al. (2013) référencent les méthodes de caractérisation de la recombinaison. Paigen and Petkov (2010) se penchent sur les caractéristiques et l'évolution des hotspots chez les mammifères. Enfin, Grey et al. (2018) et Ponting (2011) ont écrits sur le gène PRDM9 et son évolution.

I - la recombinaison chez les animaux

1. *déroulement*

Chez les animaux la reproduction sexuée implique le passage de cellules diploïdes en cellules haploïdes. Ce phénomène c'est la méiose. Durant ce cycle cellulaire un mécanisme appelé recombinaison va générer de nouvelles combinaisons d'allèles (Baudat et al. (2013), Coop and Przeworski (2007), Paigen and Petkov (2010)). En conséquence le matériel génétique transmis aux descendants ne sera pas exactement le même que celui des parents (**Figure 1)**. La recombinaison s'effectue entre deux chromosomes homologues et désigne deux mécanismes : les crossing over et les non crossing over . Les crossing over réalisent un échange réciproque et couvrent une portion importante du génome (500 à 2000 pb). Les non crossing over consistent en un transfert génétique à sens unique et à plus petite échelle (10 à 300 pb maximum) (**Figure 1)**. Ces évènements représenteraient respectivement 10 % et 90 % des évènements de recombinaisons bien qu'il ne soit pas évident d'avoir un chiffre exact .

2. ***Importance et évolution***

Il semblerait que chez la plupart des mammifères les crossing over et non crossing over aient des rôles essentiels dans la méiose : la liaison entre les chromosomes homologues et leur ségrégation par la suite (Baudat et al. (2013), Coop and Przeworski (2007), Paigen and Petkov (2010)) Il y a donc d'une part une pression positive sur le nombre de crossing over : l'aneuploïdie et donc l'infertilité qui maintient un nombre suffisant de crossing over au cours des générations . D'autre part il y a également une pression de sélection pour éviter un trop grand nombre de crossing over : la maladie de charcot en est un exemple. Le processus de recombinaison est donc très contrôlé chez un individu et à l'échelle de la population. Il en résulte un brassage génétique important pour la diversité des populations.

3. ***caractériser la recombinaison***

Il existe trois approches classiques pour étudier la recombinaison (Baudat et al. (2013), Peñalba and Wolf 2020). Tout d'abord l'analyse de pedigree qui consiste à génotyper des individus et leur descendance d'une population naturelle ou croisée en laboratoire (sur une, deux ou plusieurs générations). En comparant les différents génotypes il est possible de localiser des crossing over mais pas les non crossing over. Une autre méthode consiste à immunoprécipiter des protéines spécifiquement associées à la recombinaison comme SPO11 associé à la cassure double brin et DMC1 et MLH1 associés à la réparation des cassures double brin. Le séquençage de la chromatine associée à DMC1 permet de localiser un événement de recombinaison avec une résolution d'environ 1 kb et de détecter également les non crossing over. Enfin l'approche appelée "sperm-typing" consiste à séquencer en cellule unique des gamètes (pour l'instant des spermatozoïdes) et à comparer ces résultats avec le séquençage des cellules diploïdes du même individu (Jeffreys et al. 2001). Elle a récemment été étendue à l'analyse de génomes entiers. Cette méthode est efficace, et permet de détecter

correctement n'importe quel événement de recombinaison mais c'est également la plus coûteuse. Ces méthodes permettent d'analyser plusieurs phénotypes de la recombinaison. Le taux de recombinaison est fréquemment étudié et, dans une moindre mesure, la localisation des recombinaisons.

4. *phénotypes de la recombinaison*

L'intensité ou taux de recombinaison se focalise sur le nombre de recombinaisons le long du génome. Il y aurait des différences individuelles de ces taux et des différences spécifiques liées aux sexes. Ainsi chez les femmes les taux de recombinaisons seraient en moyenne plus important que chez les hommes (Coop and Przeworski 2007).

Chez les animaux d'élevages il existe un certain nombre de travaux sur la recombinaison et en particulier sur les taux de recombinaisons. Il est raisonnable de se demander si la sélection artificielle de ces animaux modifie le comportement de la recombinaison. Il n'existe pas pour autant de consensus sur cette question : dans une étude récente (Muñoz-Fuentes et al. 2015), des chercheurs ont noté que le taux de recombinaison n'augmente pas avec la domestication pour le mouton, la chèvre et le chien. Cependant ils soulignent également que plusieurs travaux ne vont pas dans le sens de leur recherche et montrent une augmentation du taux de recombinaison chez les plantes, les insectes et les mammifères domestiques (le cochon, le taureau, le chien , le mouton …). Deux études parues pratiquement en même temps se consacrent au taux de recombinaison et son déterminisme génétique chez des populations de moutons d'élevage (Petit et al. 2017) et de moutons sauvages (Johnston et al. 2016). Les deux travaux indiquent que les taux de recombinaisons entre les deux populations étudiées sont similaires (environ 1.5 cM/Mb). Certains gènes auraient un impact sur ce taux, notamment RNF212 qui a été identifié dans les deux études. Le taux de recombinaison aurait donc une certaine héritabilité (0.23 pour la population

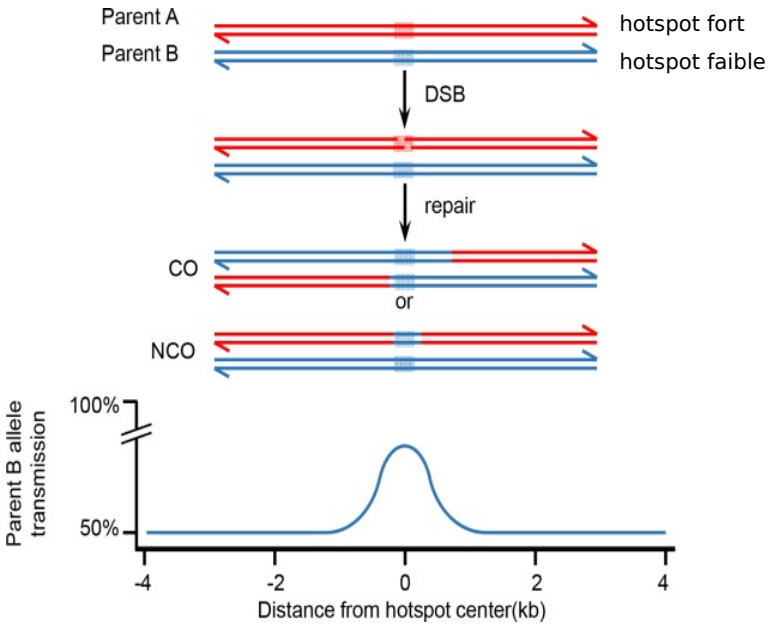


Figure 4 : érosion des hotspots *(figure provenant de Grey et al. 2018)*
Cette figure illustre le phénomène d'érosion des hotspots : le parent B présente un hotspot dit faible (moins utilisé) qui aura plus de chance d'être transmis selon le biais de conversion induit par les cassures doubles brins.

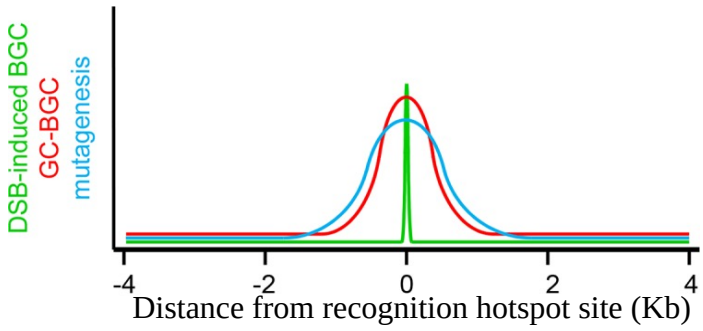


Figure 5 : conséquence de la recombinaison sur le site de reconnaissance du hotspot(*figure provenant de Grey et al. 2018*)
La réparation de la cassure double brin favorise l'apparition d'allèles GC au détriment des AC (rouge) , et de mutations (bleu) . L'allèle responsable de l'initiation de la recombinaison est sous transmis (dBCG en vert) ce qui agirait contre l'apparition de nouveaux hotspots, favoriserait les hotspots les plus faibles et conduiraient à l'extinction des hotspots dits forts

domestique (Petit et al. 2017) et 0.15 pour la population sauvage (Johnston et al. 2016). De plus Johnston et collaborateurs ont relevés une héritabilité plus importante chez les femelles et un taux de recombinaison plus important chez les mâles. C'est intéressant puisque c'est l'inverse de ce qui a été retrouvé chez l'humain (Coop and Przeworski 2007). Ainsi, une partie au moins du phénomène de recombinaison est soumis à une action génétique et un effet sexe est observé qui varie d'une espèce à l'autre.

Puisque les taux de recombinaisons ne sont pas homogènes le long du génome il reste une question : comment est dirigée la localisation de la recombinaison ? C'est un phénomène à étudier d'autant plus qu'il y a moins de travaux dessus que sur les taux de recombinaisons. Pour expliquer la localisation des recombinaisons il est nécessaire de parler des "points chauds" de recombinaison.

II - déterminisme de la localisation

1. *les points chauds*

Les points chauds de recombinaison sont des régions du génome riches en évènements de recombinaison (Paigen and Petkov 2010). Ils ont une taille d'environ 1 à 2 kb (**Figure 2)**. La caractérisation des points chauds peut se faire grâce au déséquilibre de liaison (Myers 2005). C'est à dire la fréquence à laquelle deux SNP sont retrouvés ensemble. Le déséquilibre est total si deux snp sont systématiquement retrouvés ensemble. Il s'agit d'une approche populationnelle.

2. *PRDM9*

PRDM9 est particulièrement intéressant et étudié puisque il est essentiel pour le processus de recombinaison de la plupart des mammifères. Nous allons ici nous concentrer sur le fonctionnement des espèces PRDM9 dépendantes. La structure de PRDM9 est plutôt conservée même si des zones précises sont assez variables. PRDM9 peut être divisé en trois régions avec différentes fonctions (Grey et al. (2018), Ponting (2011)) **(Figure 3)** :

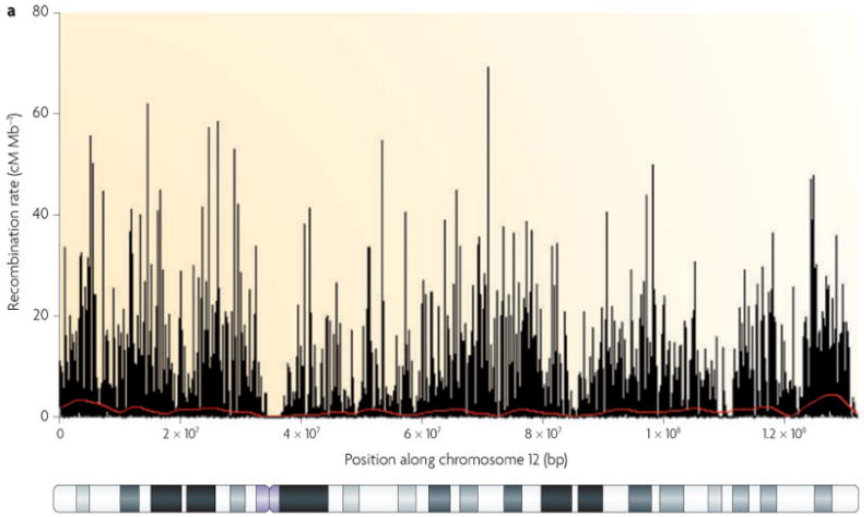


Figure 2 : taux de recombinaison sur le chromosome humain 12
(figure provenant de Paigen and Petkov 2010)
En noir le taux de recombinaison estimé, en rouge le taux de recombinaison selon la carte du projet deCODE. Les régions riches en recombinaison peuvent être des hotspots. En dessous du graphique il y a la représentation du chromosome : le centromère correspond à une région particulièrement pauvre en recombinaison.

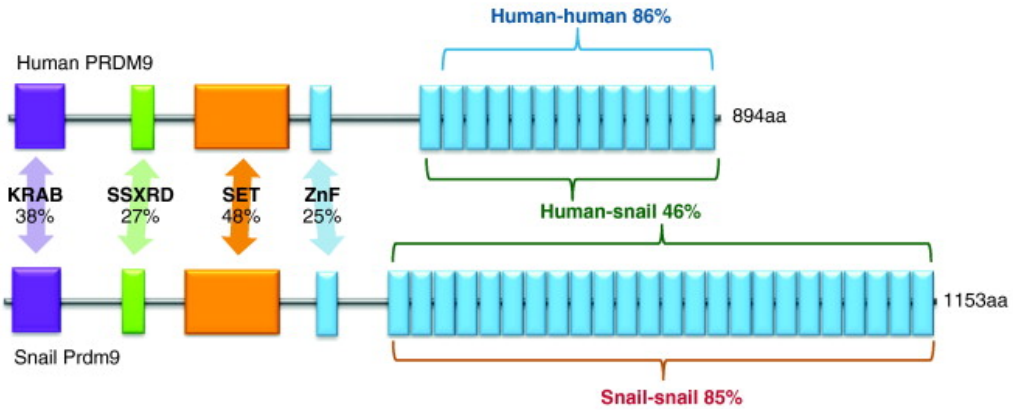


Figure 3 : structure de PRDM9, différences et similitudes entre espèces PRDM9 dépendantes *(figure provenant de Ponting)*
PRDM9 présente trois domaines essentiels : le domaine KRAB (en violet), le PRE/SET domaine (en vert et en orange) et les doigts de zinc (en bleus). Les pourcentages d'identité entre les séquence nucléotidiques sont indiqués (entre humains et escargots et au sein de l'espèce). Ils indiquent une origine ancienne des ces protéines (environ 600 millions d'années) et pourtant le pourcentage de conservation intra-espèce (86 chez l'humain et 85 chez l'escargot) indique également une évolution récente. Des duplications sont. montrées chez l'escargot.

- ***Le domaine KRAB :*** Cette partie est très conservée et permet de tracer l'histoire évolutive de PRDM9. Même chez des eucaryotes n'étant pas PRDM9 dépendant il est possible de trouver une séquence PRDM9 avec un domaine KRAB complet (dans ce cas les doigts de zinc sont souvent tronqués) (Schield et al. 2020). Il pourrait servir aux interactions protéiques concernant la recombinaison en plus d'avoir une possible utilité pour la lutte contre les rétrotransposons .
- ***Le domaine PR/SET :*** L'activité methyltransferase de PRDM9 ouvre la chromatine et c'est ce phénomène qui recrute l'ensemble des protéines, dont SPO11, responsables de la cassure double brin .
- ***Le domaine doigts de zinc :*** Cette région est la plus variable de la séquence PRDM9. Entre espèces et au sein de l'espèce il y a des nombres variables de doigts de zinc ainsi que des modifications sur des bases précises à l'intérieur de la séquence (-1, 3 et 6 chez l'humain) (Ponting 2011). Cette région est capable de reconnaître un motif d'une dizaine de paires de bases dans le génome où aura lieu la recombinaison.

3. ***évolution des points chauds chez les espèces PRDM9 dépendantes***

Les points chauds présentent un turn over important chez certains mammifères à cause des échanges fréquents entre chromosomes (Paigen and Petkov 2010). Par exemple, entre le chimpanzé et l'humain la position des points chauds n'est pas conservée malgré la proximité des deux espèces. Cela sous entend que les points chauds évoluent plus vite que la séquence du génome. Ce phénomène provient de l'érosion des points chauds causée entre autres par le biais de conversion de gène. La partie de génome transmise est celle provenant du point chaud dit "faible" (peu utilisé) **(Figure 4 & Figure 5).** Grey et collaborateurs soulignent que plus un point chaud est utilisé plus il est soumis à ce mécanisme, cela tend vers une homogénéisation des points chauds et donc une disparition de ceux qui sont "forts" (très

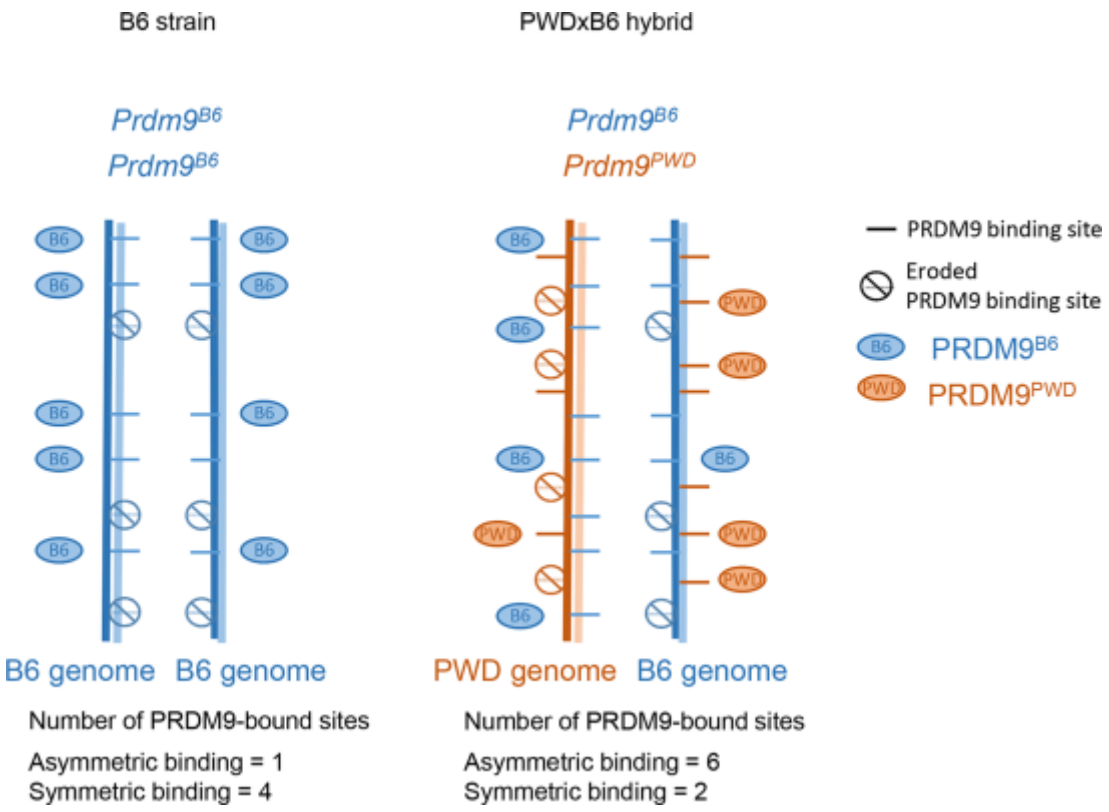


Figure 6 : stérilité hybride chez la souris *(figure provenant de Grey et al. 2018)* Si PRDM9 est hétérozygote les différentes protéines ne se lieront aux mêmes sites car certains auront été érodés (correspondant à des sites avec une forte affinité à PRDM9) d'autres non. Sur la figure PRDM9 est homozygote à gauche, les sites érodés et sains sont les mêmes donc la liaison est symétrique et peut résulter en une cassure double brin. A droite les sites érodés et sains sont assymétrique entre chromosomes homologues il ne peut pas y avoir de cassures doubles brins.

utilisés). Un modèle a été proposé selon PRDM9 évoluerait parallèlement aux points chauds (Grey et al. (2018), Paigen and Petkov (2010), Ponting (2011)).

4. *<u>évolution de PRDM9</u>*

Chez l'humain les séquences doigts de zinc de PRDM9 reconnaissent un motif 13-mer (Ponting 2011). Cette reconnaissance fait intervenir d'autres acteurs de la cassure double brin (comme SPO11) et un crossing-over peut avoir lieu (Baudat et al. 2013). Ces doigts de zinc ont un fort taux de mutation (Baudat et al. (2013), Grey et al. (2018), Ponting (2011)). Ce serait cette diversité allélique spontanée qui permettrait de compenser le phénomène d'érosion des points chauds mentionné plus haut. L'érosion des points chauds agirait donc comme une pression de sélection positive sur l'évolution et la diversification de PRDM9. Grey et collaborateurs s'y réfèrent comme à la dynamique de la "reine rouge". Ceci permettrait, entre autres, d'avoir un niveau suffisant de crossing-over. Ainsi, la non-disjonction du chromosome 21 chez l'humain aurait été reliée à un faible polymorphisme de PRDM9 (Grey et al. 2018). Cependant, quelques allèles de PRDM9 pourraient être soumis à une sélection purifiante si il se trouve en déséquilibre de liaison avec une région désavantageuse du génome. Paigen and Petkov (2010) ainsi que Grey et al. (2018) et Ponting (2011) précisent que la maladie de Charcot pour laquelle un allèle PRDM9 non A chez l'humain est corrélé à un plus faible taux de recombinaison dans la région du génome concernée et donc à un plus faible risque d'avoir la maladie . Il y a pourtant un phénomène d'homogénéisation de ce gène qui rend les allèles caractéristiques de populations (Ponting 2011).

Tout cela fait de PRDM9 le seul gène de spéciation connu chez les mammifères. Pour bien comprendre ce mécanisme il est intéressant de se pencher sur la stérilité hybride chez les souris (Baudat et al. 2013), (Grey et al. 2018) **(Figure 6).** Une souris ayant deux parents avec

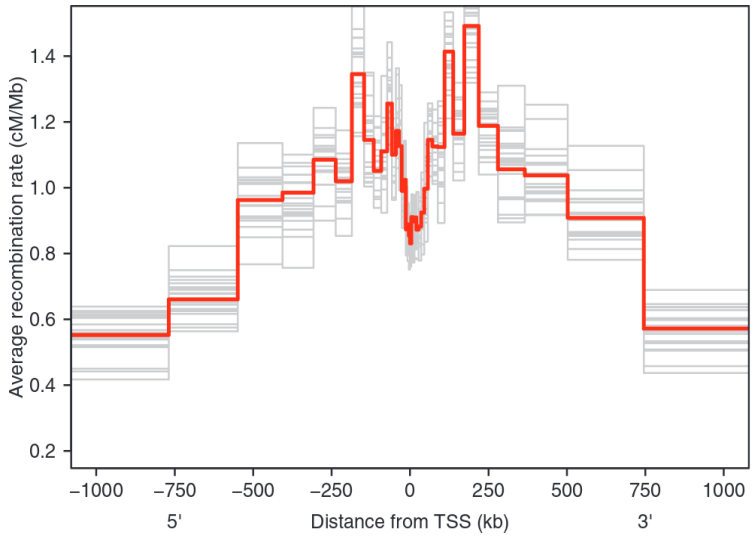


Figure 7 : distribution de la recombinaison autour d'un site de début de transcription *(figure provenant de Coop et al. 2008)* La ligne rouge représente le taux moyen de recombinaison autour d'un site de début de transcription, les lignes grises représentent l'incertitude de cette estimation. La recombinaison evite le site de début de transcription.

des génomes divergents (mus musculus domesticus et M. m. musculus) sera stérile. Chaque génome aura des sites érodés correspondant au PRDM9 qu'il porte tandis que ces mêmes sites seront intacts sur le chromosome homologue. Ainsi, il ne pourra pas y avoir de cassure double brin puisque la reconnaissance des points chauds sera asymétrique (les différents allèles ne pourront pas reconnaître les mêmes sites).

5. les différents modèles de déterminisme de la localisation de la recombinaison chez les animaux

Classiquement le déterminisme de la localisation des points chauds est associé à des genres particuliers : les oiseaux auraient une localisation des points chauds plutôt conservée ce qui ne serait pas le cas chez les mammifères (Schield et al. 2020). De plus les points chauds se situeraient essentiellement à l'intérieur des gènes et des promoteurs chez les oiseaux et ce à cause de la chromatine ouverte des ces régions qui dirigerait la recombinaison (Schield et al. 2020). A l'inverse les mammifères auraient des points chauds plutôt concentrés dans des zones intergéniques et moins dans les sites de début de transcription **(Figure 7)** (des télomères et des centromères également bien qu'il soit possible que les recombinaisons dans ces endroits ne soient simplement pas détectées) (Brick et al. (2012), Coop et al. (2008)). Ce phénomène serait dirigé par le gène PRDM9 et la reconnaissance de motifs spécifiques (Baudat et al. (2013), Grey et al. (2018), Ponting (2011)). Il est intéressant de noter que dans des expériences menées sur la souris, pour des sujets PRDM9 -/- la localisation des points chauds se déplace jusque dans l'euchromatine riche en marques H3K4me3 comme les promoteurs et les sites de début de transcription (Brick et al. 2012).

Ce sont là deux modèles principaux mais beaucoup d'études sont menées chez les mammifères et concernent l'Homme et la souris. Les travaux qui s'éloignent un peu de ces modèles révèlent que ces différentes catégories de points chauds ne sont peut être pas aussi

strictes et imperméables que ça. Une étude récente a démontré que le crotal posséderait à la fois une forme active du gène PRDM9 et un phénotype habituellement associé aux oiseaux (points chauds intragéniques) (Schield et al. 2020). Les auteurs indiquent que certains travaux montrent des phénomènes différents chez d'autres races de serpent. Il existe également des exceptions chez les mammifères. Le chien (un animal fortement soumis à une sélection artificielle) ne possède pas PRDM9 (Baudat et al. 2013), la souris possède PRDM9 mais la recombinaison sur sa région pseudo autosomale n'est pas dirigée par ce gène (Brick et al. 2012). Dans ce cadre étudier les crossing-over et le déterminisme de leur localisation chez un nouveau modèle ne peut se faire que sans à priori.

6. les études sur la localisation des points chauds

Alhawat et collaborateurs (Ahlawat et al. 2016) situent PRDM9 sur les chromosomes 1, 5 et X chez le mouton , et sur les chromosomes 1, 8 et X les chèvres. De plus ils ont remarqué des différences entre les PRDM9 portés par des chèvres et ceux portés par des moutons, cette différence pouvaient déjà s'observer entre les humains et les chimpanzés (Ponting 2011). Ils précisent également que la diversité se fait sur les doigts de zinc de PRDM9 (sur le nombre de répétitions et sur 4 positions des doigts de zinc), entre espèces et intra espèce. Ils concluent en supposant que PRDM9 activent des points chauds différents dans les deux espèces, mais n'établissent pas de lien entre PRDM9 et l'usage des points chauds.

Les travaux étudiants la localisation de la recombinaison se penchent souvent sur un phénomène appelé taux d'usage des points chauds. C'est à dire la fréquence à laquelle des évènements de recombinaison vont se retrouver dans des points chauds. Ce taux dépend de plusieurs facteurs (Pratto et al. 2014). D'une part il dépend des différences entre les populations qui servent à localiser les recombinaisons et les populations qui servent à

localiser les points chauds. D'autre part, si l'espèce est PRDM9 dépendante, le nombre d'allèles différents de PRDM9 est un facteur supplémentaire à prendre en compte .

En ce qui concerne spécifiquement la localisation des crossing-over il y a peu de recherches chez les animaux d'élevages. Sandor et al. (2012) ont trouvés des taux d'usage des points chauds chez les bovins variant de 4 à 58 % chez leurs individus (que des mâles). Ils ont également estimé l'héritabilité de ce taux à 0.21 ce qui est très proche de celle estimée chez les humains qui est de 0.22 (Coop et al. 2008). Les travaux de Sandor et al. (2012) et ceux de Ma et al. (2015) mettent en évidence une étude d'association qui permet de supposer ce que taux d'usage est sans doute dirigé par PRDM9 chez les bovins. Ils précisent que l'origine de ce phénotype est moins polygénique que celle du taux de recombinaison. Leurs analyses d'association ont situé PRDM9 sur le chromosome 1 et deux copies sur le X. Ces deux études ont été réalisées avec des puces de respectivement 50K et 60K pour la détection des points chauds et de 3 à 50K et de 60K pour la détection des crossing over pour étudier le taux de recombinaison et l'usage des points chauds. Un points chauds étant une petite région (de 1 à 2 kb) et le crossing over un évènement très localisé dans le génome, ces résolutions peuvent donner un taux d'usage ne correspondant pas totalement à la réalité. En revanche ils ont des populations d'au moins 10 000 individus, les résultats sont donc robustes. Enfin, ces travaux se basent sur un taux d'usage des points chauds dit brut (c'est à dire un pourcentage de crossing-over tombant dans des régions points chauds sans correction).

Dans une l'étude sur les humains de Coop et al. (2008) , le taux d'usage des points chauds est corrigé pour prendre en compte la probabilité de recouvrement aléatoire entre crossing-overs et points chauds (cf. ci-dessous). De plus ils utilisent une puce 500K pour caractériser les recombinaisons de manière précise. Les points chauds qu'ils utilisent proviennent d'une précédente étude (Myers 2005) caractérisés avec une précision sur la base

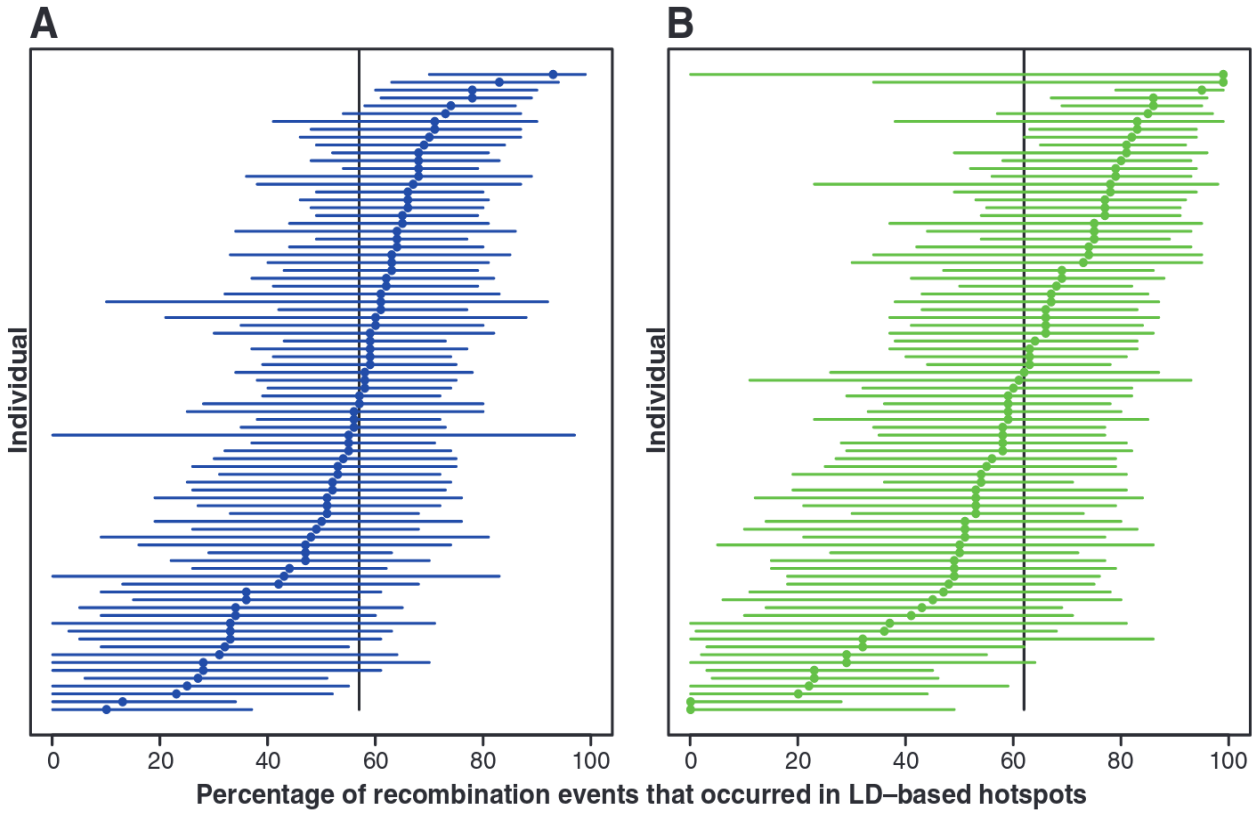


Figure 8 : proportion de cross-over apparaissant dans des hotspots par individu *(figure provenant de Coop et al. 2008)*

Le poucentage de cross over estimé par maximum de vraisemblance qui apparaissent dans des hotspots pour chaque individu (les femelles (A) et les mâles in(B) est indiqué par un cercle. Les intervalles de confiance sont représentés par la longueur des barres horizontales. Les lignes noires verticales est le maximum de vraisemblance globale.

de 1.6 million de marqueurs génotypés dans des populations africaines, américaines et européennes. Chez les individus de l'étude ils estiment des taux de recouvrement des points chauds variant de 0 à 100 % entre individus sans trouver de différences entre mâles et femelles **(Figure 8)**.

III - objectifs

Ainsi il n'y pas beaucoup d'études sur l'usage des points chauds puisque ce phénotype est difficile à caractériser avec précision. Par extension la corrélation entre ce phénomène et PRDM9 est également moins étudié que les associations avec le phénotype d'intensité. De plus les études s'y intéressant sont effectuées essentiellement chez les humains mais nous avons mentionné que les mécanismes impliqués dans l'usage des points chauds chez l'humain ne sont pas généralisables à toutes les espèces.

Cela pose quatre questions :

- Y a t-il un chevauchement significatif entre crossing-over et points chauds de différentes populations de moutons?
- Si oui, y a t-il des différence individuelles entre ces taux d'usages ?
- Serait-ce dû à une origine génétique ?
- Si oui, peut on l'imputer à PRDM9 ?

La première étape consiste à détecter les crossing over avec une bonne résolution. C'est à dire trouver des crossing over compris dans des intervalles suffisamment petits pour que la suite des analyses soit précise. Dans un deuxième temps ces crossing over serviront à estimer des taux d'usage de points chauds ou la fréquence de chevauchement des points chauds par des crossing over. Plusieurs taux d'usages seront cherchés : global, par sexe et par individu. Finalement une analyse d'association génétique fournira les renseignements sur les liens entre ces phénotypes (les taux d'usage) et la génétique.

2) RÉALISATIONS

I - matériels et méthodes

1. Les données

Les données utilisées dans le cadre ce travail proviennent du génotypage réalisé avec des puces 600 K de 375 moutons de la race Romane. La puce 600 K permet un balayage précis du génome puisqu'elle permet d'identifier un marqueur tous les 5 Kb environ. Enfin, la race Romane est intéressante à étudier, car étant plutôt récente (créée autours des années 1970), elle représente l'hybridation de deux races éloignées (la Romanov et la Berrichon du Cher) donc un certain polymorphisme. Après un nettoyage des données 7 individus ont été écartés (un individu dupliqué, un génotype trop peu informatif et 5 erreurs d'affiliation). En plus de ces données de génotypages, un jeu de données de 45358 points chauds, provenant d'un travail antérieur (Petit et al., 2017) réalisé sur une population de moutons Lacaunes, a été ajouté à cette étude.

2. détection des crossing over

Cette étape est cruciale pour les analyses qui suivent. Ainsi deux outils ont été utilisés pour la détection des crossing over : LinkPHASE et DuoHMM.

a) LinkPHASE

LinkPHASE est un logiciel servant, entre autres, à détecter des évènements de crossing-over (Druet and Georges 2015). Pour cela il "phase" les génotypes, c'est à dire qu'il va assigner à chaque allèle d'un SNP hétérozygote d'un individu une origine maternelle ou paternelle (Druet and Georges 2015). Dans le cas où au moins l'un des parents est homozygote, l'attribution des origines est non ambiguë. Dans le cas où les deux parents sont hétérozygotes, le logiciel va phaser les SNP en se basant sur ceux déjà phasés (Druet and Georges 2015) et en maximisant la vraisemblance des génotypes observés qui dépend des probabilités de recombinaison. Ce processus est la base de LinkPHASE mais la dernière

implémentation (Druet and Georges 2015) de ce logiciel a ajouté un modèle de Markov caché qui améliore ses performances dans le cas où les familles sont incomplètes (demi-frères, demi-soeurs). En revanche il est impossible de phaser des individus avec deux descendants ou moins sauf s'ils ont eux mêmes des parents. Dans mon travail, j'ai repris les scripts disponibles de l'étude de Petit et al. (2017) pour analyser les données.

b) DuoHMM

DuoHMM se compose de deux parties utilisant chacune des informations de nature différentes. La première partie consiste à trouver pour un individu son haplotype le plus probable sans utiliser l'information de pedigree : le principe est de tirer parti de la conservation des associations entre allèles proches d'une population (c'est le déséquilibre de liaison) (Delaneau et al. 2012). Dans un second temps, le logiciel va confronter cette reconstruction initiale aux informations de pedigree et corriger d'éventuelles erreurs (O'Connell et al. 2014). Ainsi, par rapport à LinkPHASE, il utilise une information supplémentaire : le déséquilibre de liaison entre marqueurs. Ce logiciel a de plus longs temps de calculs mais permet de mieux exploiter les données que LinkPHASE : il est possible d'obtenir la phase d'un individu avec un seul descendant. Ce logiciel nécessite une carte génétique en entrée : nous avons utilisé une carte provenant d'un travail antérieur (Petit et al., 2017) provenant d'une puce de 60 000 SNPs.

3. calcul des taux d'usage des points chauds de recombinaisons

On appelle taux d'usage des points chauds (noté alpha) se caractérise par la propension qu'ont des évènements de recombinaisons à se situer dans des points chauds de recombinaisons. Ici deux approches pour estimer alpha sont proposées : un taux estimé de manière naïve (TN) et un taux estimé grâce à un maximum de vraisemblance (TML).

a) Approche du Taux Naïf

Pour chaque individu le taux d'usage des points chauds peut être estimé naïvement par la proportion de crossing over chevauchant des points chauds. En utilisant pybedtools (paramètres : overlap -u) les intervalles contenant un crossing over sont confrontés à un fichier réunissant les points chauds de la population Lacaune. Un chevauchement est compté qu'il soit partiel ou complet, finalement le résultat présenté est le rapport entre le nombre de chevauchements et le nombre d'intervalles total. Les intervalles dans lesquelles des crossing over ont été détectés ont été répartis en fonction de leur taille (5 kb, 10 kb, 20 kb, 30 kb et ainsi de suite de 10 kb en 10 kb jusqu'à 1 Mb). Cette approche ne prend pas en compte la variabilité des tailles des crossing over ni leur environnement. Pour le premier point : plus un intervalle de crossing over est grand plus il a de chance de chevaucher un point chaud. Le second point est important si un crossing over est détecté dans une région riche en points chauds. Dans ce cas il aura plus de chance d'en chevaucher un. Pour pallier ces éventuels effets la méthode de Coop et al. (2008) a été utilisée.

b) Approche du Taux de Maximum de Vraisemblance

Pour chaque individu (puis pour l'ensemble des individus, des mâles et des femelles) alpha est estimé : c'est la proportion d'événements de crossing over (compris dans 30 kb) qui apparaissent dans des points chauds. Pour ce faire chaque crossing over (*r*) est testée :

P(*r* chevauche un points chauds) = TML + (1 - TML) * *P*(*r* chevauche un point chaud par hasard)

La probabilité que le crossing over chevauche un point chaud dépend de TML et de son environnement. Le premier couvre les possibilités entre 0 et 1 avec un pas de 0.01. L'environnement correspond aux termes : (1- TML) * *P*(*r* chevauche un points chauds par hasard). *P*(*r* chevauche un points chauds par hasard) correspond à une proportion de chevauchement de 100 déplacements aléatoires du crossing over d'une distance choisie au

hasard dans une distribution normale de moyenne 0 et d'écart type de 200 kb. Ainsi si la région entourant le crossing over est riche en point chaud la probabilité que le crossing over chevauche un point chaud va augmenter. Par la suite la vraisemblance pour un TML donné est calculée:

L(TML) = δr*P(r chevauche un points chauds) + (1-δr)(1-P(r chevauche un points chauds))

δr vaut 0 si le crossing over ne chevauche pas un point chaud et 1 si il en chevauche 1. Si le crossing over chevauche un point chaud la vraisemblance sera la probabilité calculé précédemment. Le TML est propre à un individu Ainsi tous ces crossing over auront le même TML. Le TML retenu pour un individu est donc celui qui maximise la vraisemblance L(TML). L'intervalle de confiance correspond aux TML compris dans ce maximum L(TML) - 2.

c) <u>estimation de la pertinence des taux d'usage</u>

Pour le TN deux tests ont été réalisé. Le premier test cherchait à explorer la différence de résolution (de taille) entre les crossing over des mâles et ceux des femelles. Il s'agit du test non paramétrique de kolmogorov smirnov qui a été réalisé sur la distribution de la taille des intervalles entre mâles et femelles (est-ce que, si il y a une différence de chevauchement entre ces populations, cette différence est liée à une différence de résolution entre mâles et femelles ?). Un second un test a été réalisé pour avoir une idée du taux qui serait obtenu par hasard. Pour chaque individu ses intervalles ont été artificiellement déplacés d'une distance choisie au hasard dans une distribution normale (de moyenne 0 et d'écart-type 200 kb puis 500 kb et finalement 1 MB). Après ces déplacements un nouveau taux est calculé. Cette opération est répétée 100 fois afin de savoir quelle est la probabilité de trouver par hasard le taux réellement estimé.

Pour le TML afin de tester si il existe une différence de taux d'usage entre groupes (sexe, individus …). Nous avons utilisé le test du rapport de vraisemblance calculé comme la différence de log vraisemblance pour un groupe ayant un unique taux et un groupe avec des taux par niveaux. Une p-valeur peut être associée à ce rapport de vraisemblance en permutant les crossing over au sein d'un groupe. C'est à dire en redistribuant aléatoirement les intervalles entre niveaux. A la suite de cette redistribution un nouveau rapport est calculé. Ces opérations ont été répétées 10 000 fois pour Linkphase et 500 fois pour DuoHMM. La p valeur est la proportion du rapport de permutation égalant ou dépassant le véritable rapport. Il est intéressant de noter que cette distribution de rapport provenant des permutations devrait suivre une loi du khi2 selon le théorème de Wilks.

4. GWAS

Les études sur les ruminants mentionnées plus haut (Ma et al. 2015 ; Petit et al. 2017 ; Sandor et al. 2012) , réalisent des études d'association pour relier les phénotypes qu'ils ont détectés à un génotype particulier. Dans les trois cas ils utilisent un modèle linéaire mixte univarié. Pour calculer ce modèle Petit et al (2017) et Ma et al (2015) utilisent des logiciels existant : gemma et bimbam (pour imputer des génotypes manquants) (Petit et al. 2017) et MMAP (Ma et al. 2015). De l'autre côté, Sandor et al (2012) ont implémenté eux mêmes leur modèle grâce à un test de ratio de vraisemblance.

Nous avons réalisé l'analyse d'association à l'aide du logiciel Gemma (Zhou and Stephens 2012) sur un modèle linéaire mixte univarié : $y = W\alpha+x\beta+u+e$ où y est un vecteur de phénotypes (ici le TML calculé ci dessus) , W correspond aux covariables (ici le sexe de l'individu) avec les effets α , x le vecteur des génotypes à un marqueur (β l'effet du génotype), u un vecteur des effets polygéniques ($u \sim N(0,G\sigma 2g)$ où G est la matrice d'apparentement)) et e les erreurs résiduelles. A chaque SNP testé le logiciel regarde si β (l'effet du génotype)

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

Table 1 : résultats des seuils de significativité pour m échantillons
(table provenant de Storey and Tibshirani 2003)
F est le nombre de faux positifs, T celui de vrais positifs. Ces deux nombre donnent T qui est l'ensemble des éléments déclarés significatifs. De plus m0 est le nombre de résultats réellement nuls et m1 le nombre de résultats relevant réellement de l'hypothèse alternative. Ces mesures sont utilisées pour estimer globalement des seuils de p-valeur.

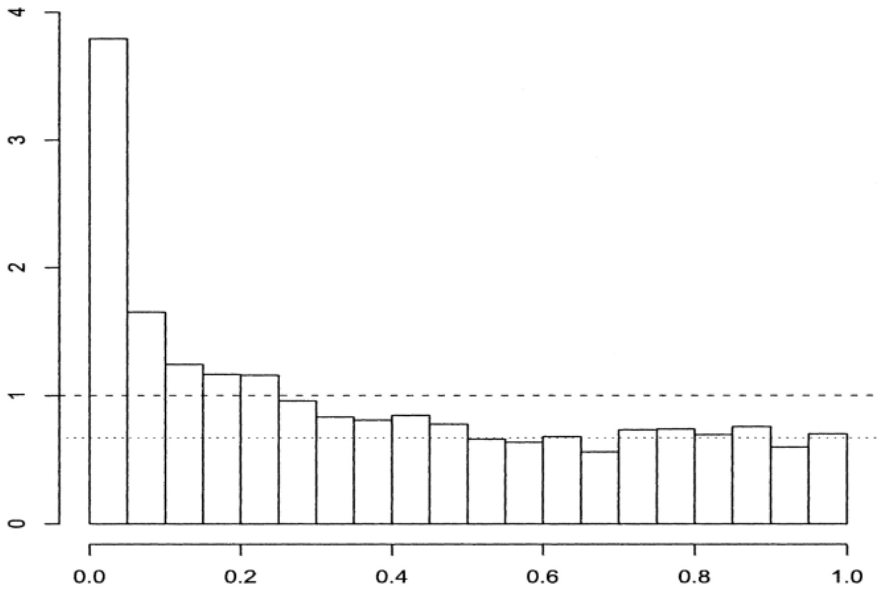


Figure 9 : histogramme de densité de p-valeur
(figure provenant de Storey and Tibshirani 2003)
Il s'agit d'une représentation graphique de la table1. La ligne avec des tirets serait l'histogramme de densité si les p-valeurs appartenaient toutes à l'hypothèse nulles. La ligne en pointillé est l'estimation de la proportion des p-valeurs appartenant réellement à l'hypothèse nulle autrement dit m0. Donc m1 suit la densité de l'histogramme jusqu'à ce que l'on distingue plus m1 et m0.

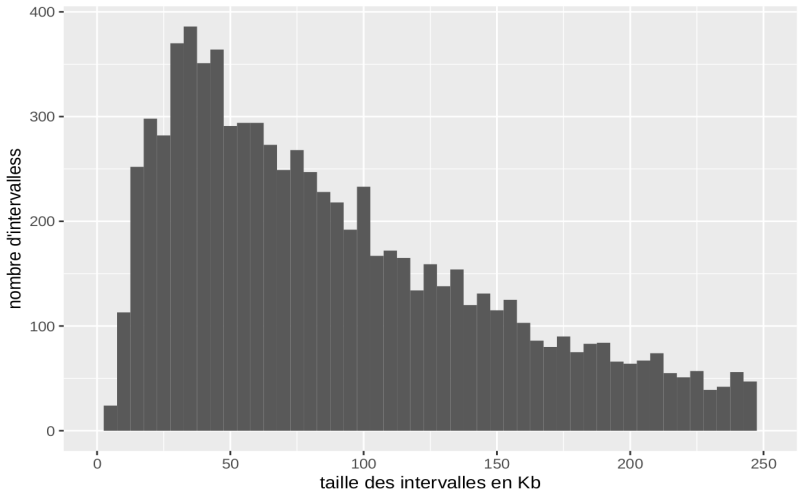


Figure 10 : distribution de la taille des intervalles pour Linphase

Distribution de la taille des intervalles. La fonction ecdf nous apprend que 70% des intervalles font moins de 230 Kb. C'est pourquoi le graphique s'arrête à une taille de 250 Kb.

résolution	overlap	P overlap	P number	M overlap	M number	chance 200 kb	chance 500 kb	chance 1 Mb	K alpha	K pvalue
< 5 Kb	100.00	100.00	7	0	0	12.43	9.86	10.14	0	0
< 10 Kb	85.51	87.50	64	60.00	5	13.83	13.36	12.13	0.38	0.43
< 20 Kb	76.09	77.27	484	62.79	43	21.55	19.13	18.60	0.15	0.30
< 30 Kb	77.27	78.52	1057	62.07	87	26.52	24.10	22.67	0.07	0.80
< 40 Kb	77.40	78.21	1767	66.14	127	31.31	28.25	26.91	0.10	0.18
< 50 Kb	78.31	78.81	2416	71.18	170	34.86	31.55	30.13	0.08	0.21
< 60 Kb	77.97	78.46	2976	70.65	201	37.65	34.18	32.59	0.09	0.12
< 70 Kb	78.58	79.00	3491	72.00	225	40.31	36.63	34.96	0.09	0.05
< 80 Kb	79.03	79.39	3984	73.20	250	42.61	38.87	37.24	0.09	0.04
< 90 Kb	79.72	80.02	4440	74.72	265	44.77	41.02	39.46	0.10	0.01
< 1Mb	80.31	80.62	4834	75.00	280	46.57	42.86	40.88	0.11	0.00

Table 2 : taux d'usage des hotspots par résolution pour linkphase

Trois taux d'usage sont présentés : le taux d'usage global, paternel et maternel. Le nombre de crossing over pour les pères et les mères a été ajouté. Les colonnes chance sont les taux d'usage trouvés par hasard pour des perturbations de 200, 500 et 1000 Kb. Les deux dernières colonnes rendent compte du test de Kolmogorov Smirnov avec le résultat alpha du test et la p-valeur associée

est significativement différent de zéro et y associe une p-valeur. Dans le cas de ce travail il est préférable de minimiser les faux positifs afin de trouver des régions du génome susceptibles d'êtres intéressantes (avec une association).Pour fixer la p-valeur le seuil a été fixée d'après le critère de FDR (False discovery rate) en admettant un FDR de 0.05 (Storey and Tibshirani 2003) **(Table.1 ; Figure 9)**. Le FDR est l'espérance du nombre de faux positifs sur l'ensemble des tests déclarés significatifs :

$$\mathrm{FDR} = \mathrm{E}\left[\frac{F}{F+T}\right] = \mathrm{E}\left[\frac{F}{S}\right]$$

Pour ce faire nous avons utilisé le calcul de la q-valeur (équivalente du FDR) (Storey and Tibshirani 2003) qui fixe une p-valeur seuil pour laquelle le FDR est de 0.05. Pour LinkPHASE 1000 permutations ont été réalisées sur les phénotypes (sans effet sexe) afin de déterminer la chance qu'il y avait de trouver par hasard les SNP significatifs. Après chaque permutation Gemma est relancé.

II - résultats

1. *Analyse du taux d'usage des points chauds avec LinkPHASE*

a) *taux d'usage naïf*

L'approche LinkPHASE a permis d'identifier 11095 crossovers dans 43 individus (36 mâles et 7 femelles) et 308 méioses. Parmi ces crossing over 5114 font moins de 100 Kb et 1144 moins de 30 Kb **(Figure 10).** Le calcul du TN nous indique que pour des intervals inférieurs à 30 Kb le taux d'usage moyen des points chauds est de 77 % **(table 2)**. En permutant la localisation des points chauds nous pouvons estimer la distribution du taux d'usage sous l'hypothèse nulle (TN = 0) **(annexe 1)**. Aucun taux permuté n'a de valeurs aussi fortes que le taux des données non permutées. La moyenne de ces taux d'usage est de seulement 26 %, autrement dit c'est le taux d'usage qui serait obtenu par hasard. La différence entre le taux d'usage des mâles et des femelles est significative **(table 2)**. Le

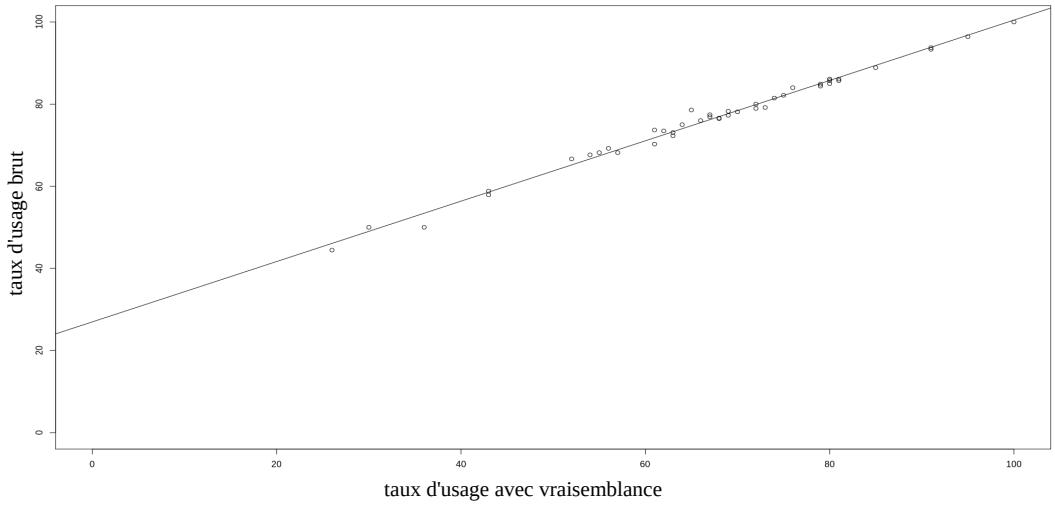


Figure 11 : taux d'usage corrigé alpha en fonction du taux d'usage brut pour Linkphase
Le taux d'usage corrigé est en en abscisse et le brut en ordonnée. Le nuage de point suit une droite qui nous indique que le taux brut surestime le taux corrigé surtout pour les valeurs faibles.

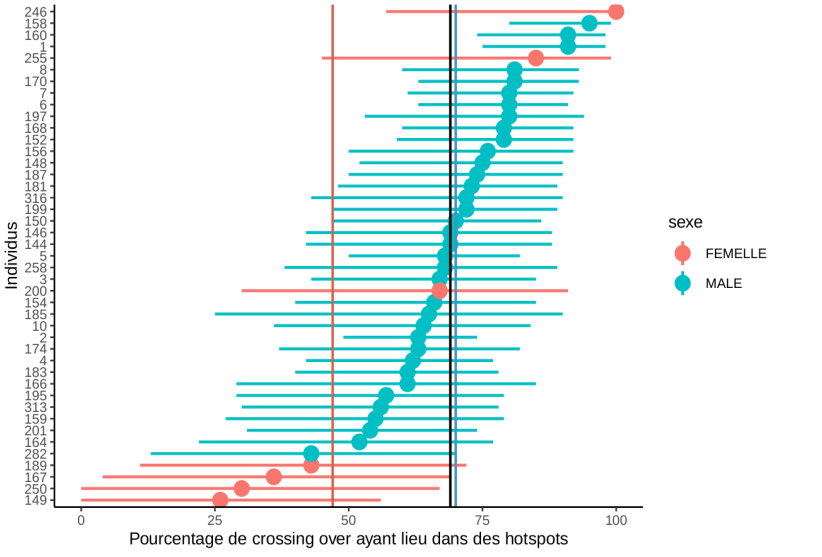


Figure 12 : pourcentage de cross-over apparaissant dans des hotspots par individu pour linkphase
Pourcentage de cross-overs détectés dans des intervalles de 30 kb qui ont lieu dans des hotspots basés sur une inférence de déséquilibre de liaison. Ces pourcentages sont calculés pour chaque individu dont les identifiants sont en ordonnée. L'estimation du maximum de vraisemblance est représentée par les points (bleus pour les mâles, rouges pour les femelles) tandis que les barres horizontales indiquent les intervalles de confiance (95 %). Les lignes verticales représentent l'estimation du maximum de vraisemblance pour l'ensemble des individus (noir), l'ensemble des femelles (rouge) et l'ensemble des mâles (bleu).

h0	LRT	ddl	p-value	pchisq
α_unique = α_sexes	11.26	1	0.001200	0.0007919521
α_unique_mâle = α_indiv_mâles	44.66	35	0.167100	0.1330208
α_unique_femelle = α_indiv_femelles	12.88	6	0.069000	0.0434117
α_sexes = α_indiv_sexes	57.54	41	0.073500	0.0463379

Table 3 : likelihood ratio pour les hypothèses sexes
La première colonne fait état des différentes hypothèses nulles : est-ce qu'un taux d'usage unique est égal aux taux d'usage pour les mâles et les femelles? Est-ce que le taux d'usage d'un sexe est égale aux taux d'usage des individus ce sexe ? Enfin la somme des deux dernières hypothèses. L'avant dernière colonne est la p-valeur du likelihood ratio calculé avec 1000 permutations d'intervalles. La dernière colonne est la p-valeur de la distribution du khi2 de ces permutations.

nombre de crossing over détectés est différent pour ces deux catégories cependant le test de Kolmogorov-Smirnov nous indique que la distribution de la taille des intervalles (ceux inférieurs à 30 Kb) ne diffère pas significativement en mâles et femelles **(table 2).** Cela suggère que les différences observées ne proviennent pas d'une différence de résolution.

b) taux d'usage estimé par maximum de vraisemblance

Nous avons ensuite utilisé une approche par maximum de vraisemblance sur les intervals inférieurs à 30 Kb qui permet de pallier les effets de la variation de la taille des intervals et de la concentration en points chauds le long du génome (voir matériels et méthodes). La comparaison du TN et du TML montre que le premier taux présente des valeurs plus fortes que le second taux **(Figure 11)**.

Cette nouvelle estimation donnent des taux variables entre individus (entre 25 et 100 % ; CI : 19 - 67 %) **(Figure 12)**. Il est de 69 % en moyenne. Ce qui cohérent avec nos résultats bruts, est la différence entre mâles (taux : 70 %) et femelles (taux : 47%). Cependant les femelles ont des taux plus variables (entre 25 % et 100%) que les mâles qui forment un groupe plus homogène (entre 42 et 95%). Nous avons testé la significativité de ces différences (voir matériels et méthodes) avec un rapport pour chaque hypothèse et une p-valeur associée **(table 3)**. L'hypothèse selon laquelle il y a une différence d'usage selon le sexe se vérifie (*p=0.0012*). La différence d'usage des individus au sein des femelles est suggestive (*p=0.069)* et non-significative chez les mâles (*p=0.167*). Nous pouvons supposer que les différences individuelles observées résultent en partie de ces différences de sexe. Ces rapport ont été calculés en permutant les crossing over entre individus. Les distributions de ces rapports pour les différentes hypothèses suivent une loi du khi2 **(annexe 2)**. D'ailleurs la p-valeur des rapports sont très similaires à celles des distributions du khi2. Cela indique que la convergence des test du rapport de vraisemblance vers une loi du khi2 sont respectées ici

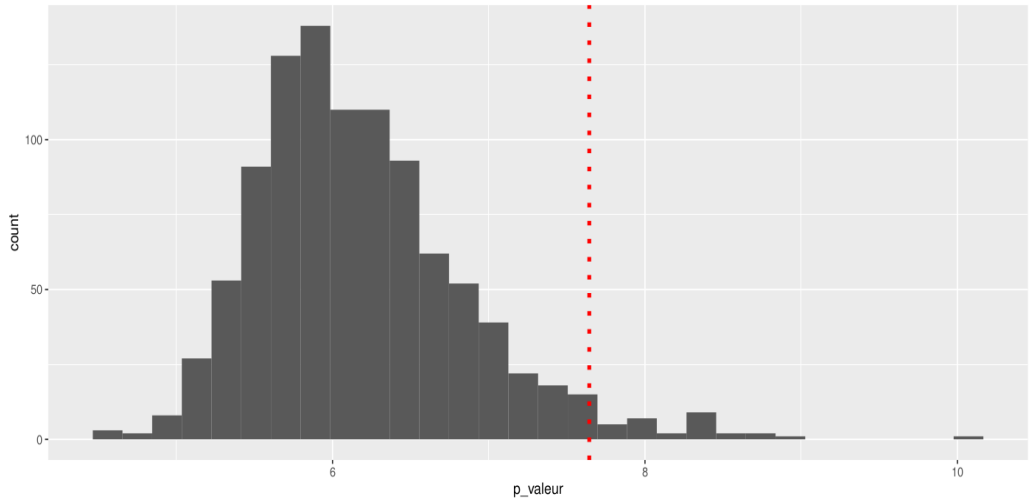


Figure 13 : probabilité de trouver un SNP significatif qui est un faux positif
Il s'agit de la distribution des log10 des p-valeurs minimales de 10 000 permutations de phénotypes. Ceci a été réalisé pour l'analyse des crossing over de tous les individus de Linkphase. Seuls les autosomes sont conernés. La ligne rouge est la p-valeur trouvée significative sur le manhattan plot. La proportion de p-valeurs au delà est de 0.032.

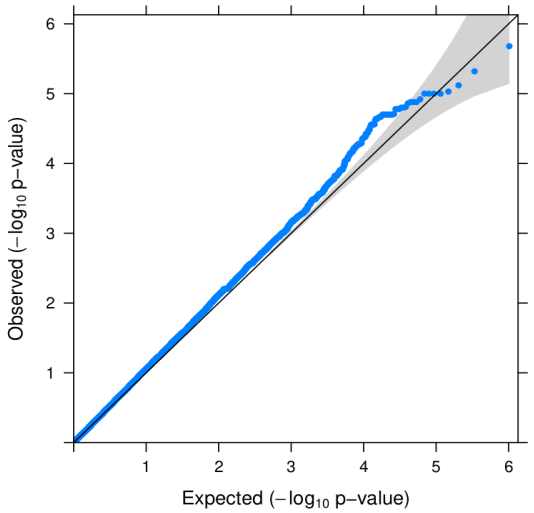


Figure 14 : ajustement des distributions des p-valeurs des LRT à un modèle théorique pour Linkphase
Il s'agit des p-valeur du test LRT pour chaque SNP calculées par Gemma. L'analyse est réalisée sur 36 mâles et 27 chromosomes En abscisse il s'agit des p-valeurs théoriques et en ordonnée les p-valeurs observées. La droite est la droite d'ajustement avec l'écart autorisé en gris. Pour avoir des données pertinentes il faut que le maximum de points suive la droite ou soit dans l'écart autorisé. Si quelques valeurs maximales se détachent de cette zone elles sont considérées significatives.

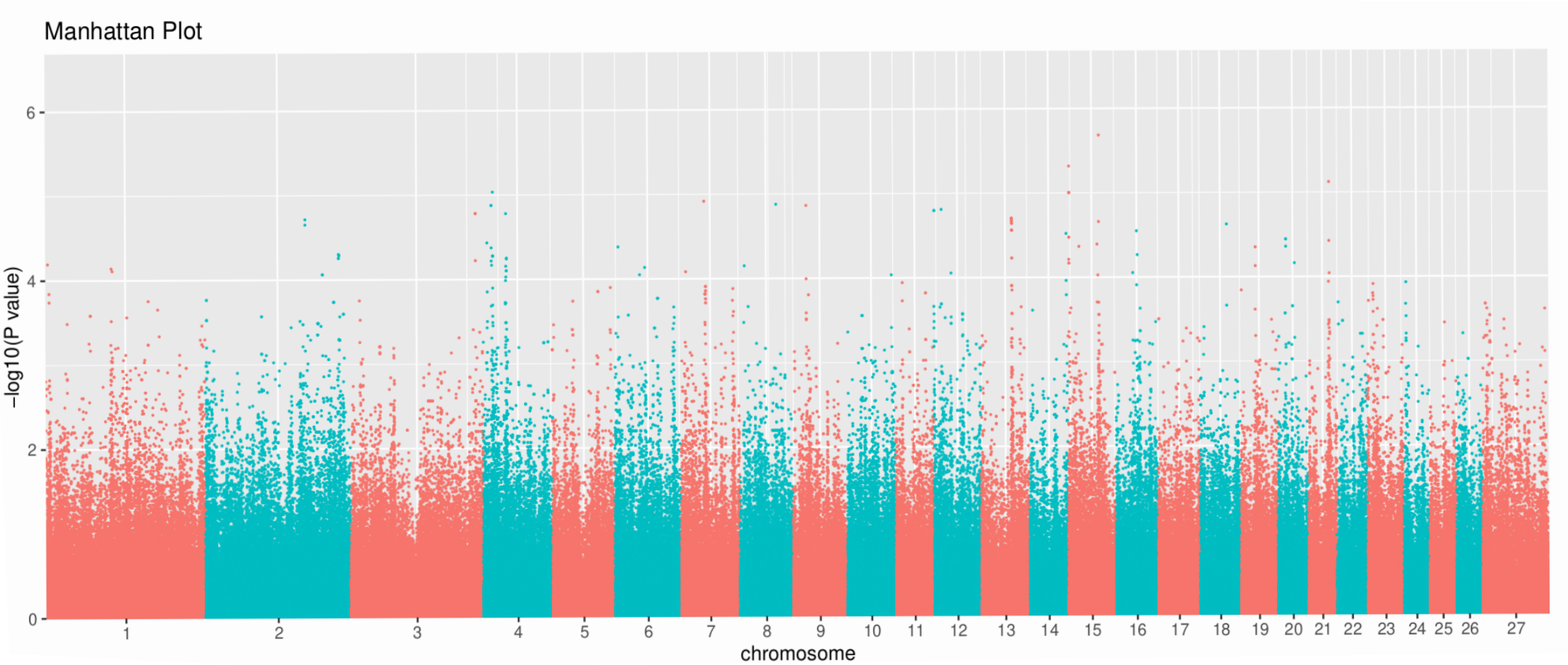


Figure 15 : manhattan plot des p-valeurs des effets des SNP pour Linkphase
En abscisse : les 26 autosomes et le chromosome X (27) et les positions génétiques
En ordonnée : les log10 des p-valeurs LRT des effets des SNP calculés par Gemma
La ligne rouge est la valeur seuil fixée avec un FDR de 0.05. 36 mâles ont été pris en compte

c) GWAS

Le logiciel utilisé propose deux méthodes de calcul des p-valeurs des effets génétiques. Le test du rapport de vraisemblance (LRT) et le test de wald. Ces deux méthodes donnent des résultats similaires **(annexe 3)** nous allons donc nous concentrer sur une seule. En l'occurrence le LRT. Gemma propose également de calculer deux matrices d'apparentement : une standardisée et une centrée **(annexe 4)**. Les auteurs (Zhou and Stephens 2012) précisent que pour les organismes avec de grand génomes comme les animaux la différence n'est pas importante **(annexe 5)**. Cependant la matrice centrée devrait donner un poids moins important aux SNP à fréquence faibles. Nous allons nous concentrer sur les résultats des matrices centrées.

Nous avons regarder un première fois les résultats de gemma sur l'ensemble des individus et sans le chromosome X pour des phénotypes de TML. Sur ce modèle un SNP avec une p-valeur de 0.0108 est significatif avec un FDR de 0.05 **(annexe 6).** Nous avons effectué 1000 permutations des phénotypes et relancé Gemma pour chacune. La distribution de la p-valeur minimale pour chaque permutation nous indique que la chance de tomber sur un faux positif est de 0.032 **(Figure 13).** Ce test nous permet de constater que notre analyse est plutôt robuste aux faux positifs.

Il y a peu de femelles dans cette analyse et elles ont moins de descendants que les mâles (36 individus). C'est pourquoi pour l'analyse d'association nous avons décidé de les retirer du jeu de données. Nous avons appliqué Gemma à nos phénotypes de TML. L'ajustement des distributions des p-valeurs à un modèle théorique nous renseigne sur la pertinence des données **(Figure 14)**.Le modèle utilisé semble avoir des p-valeurs ajustées **(Figure 14).** Pour ce modèle nous avons placé la q-valeur ou FDR à 0.05 **(Figure 15)**. Aucun SNP ne ressort.

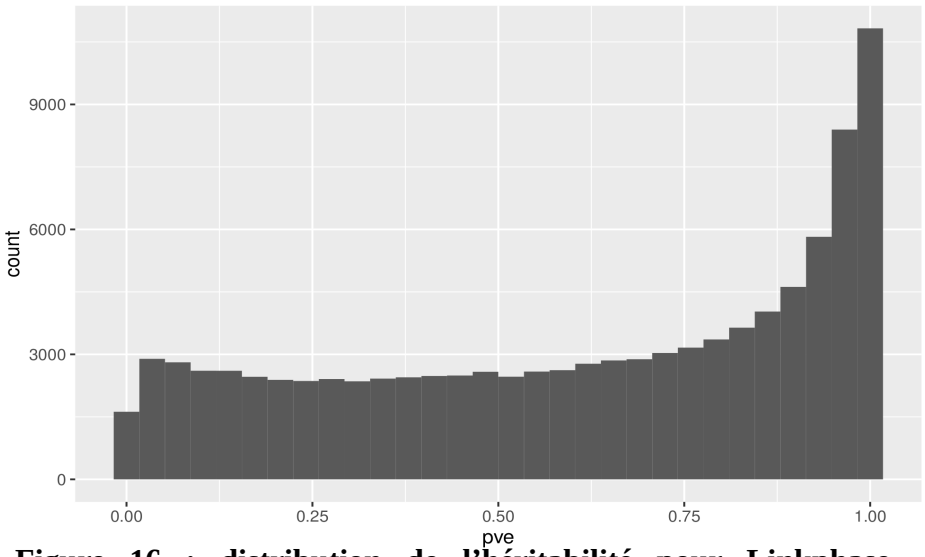


Figure 16 : distribution de l'héritabilité pour Linkphase
Distribution de l'héritabilité trouvée le modèle bayésien mixte de Gemma.
En abscisse : pve ou estimation de l'héritabilité
Avec la fonction ecdf on trouve que plus de 70 % des héritabilités se situent au dessus de 0.25

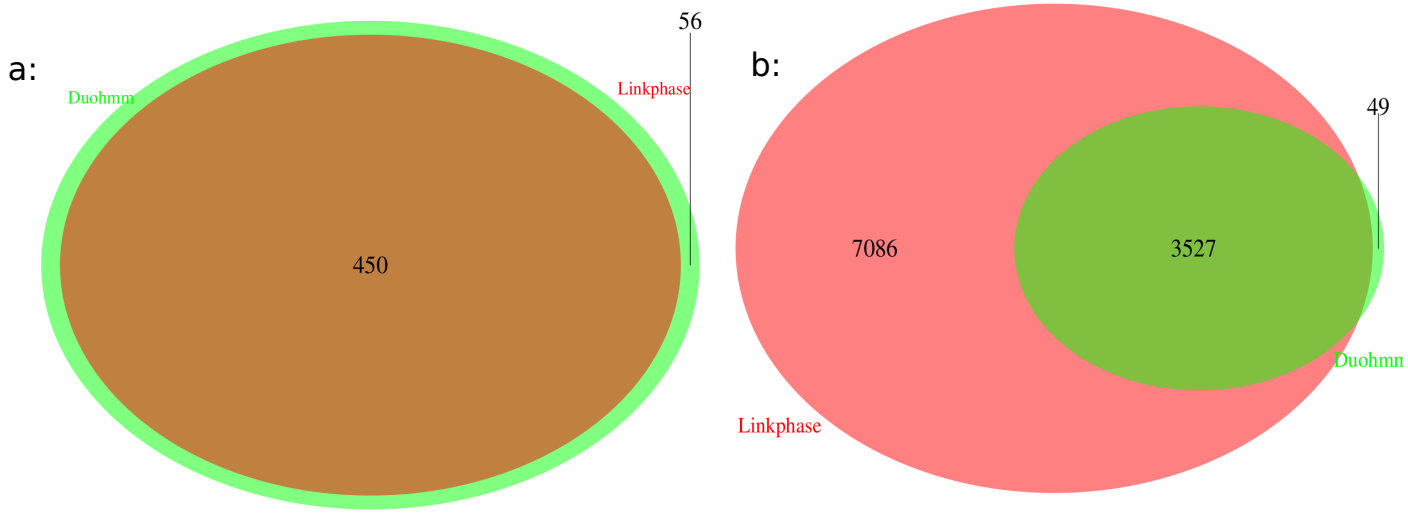


Figure 17 : proportion des taux d'usage détectés par linkphase retrouvés par duohmm
Les taux d'usages retrouvés par duohmm (en vert) initialement detectés par linkphase (rouge) pour les individus partagés pour ces deux logiciels. Il y a une représentation pour les femelles (a) et les mâles (b).

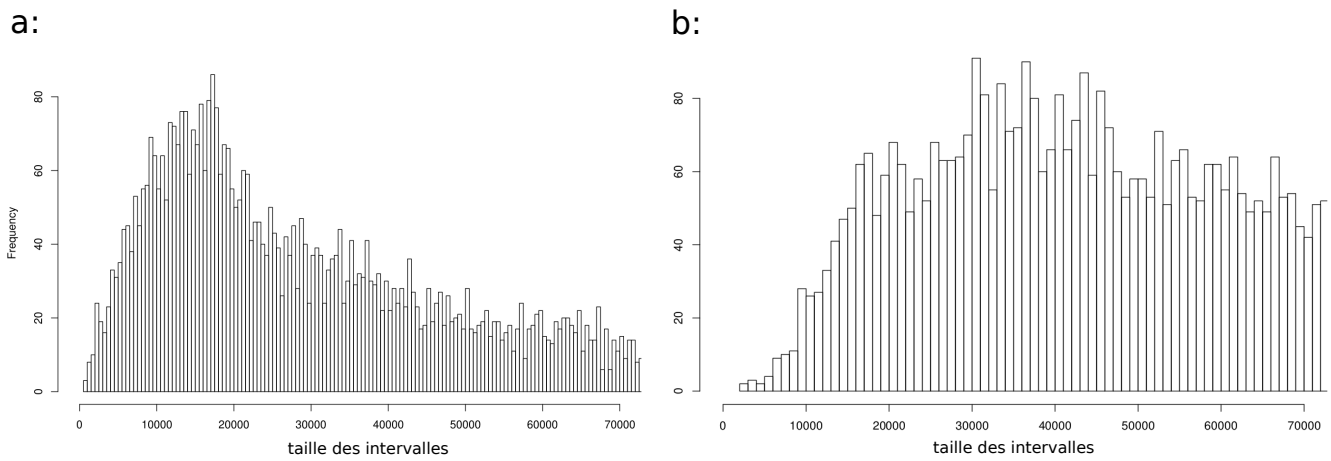


Figure 18 : distribution de la taille des intervalles

a : distribution des la taille des intervalles pour duohmm (probabilité de rebombinaison > 0.9)

b: distribution de la taille des intervalles pour linkphase

résolution	overlap	P overlap	P number	M overlap	M number	chance 200 kb	chance 500 kb	chance 1 Mb	K alpha	K pvalue
< 5 Kb	66.67	85.71	21	51.85	27	11.4	1023	9.88	0.23	0.49
< 10 Kb	74.55	83.18	107	66.3.7	113	14.43	13.09	11.90	0.23	0.01
< 20 Kb	68.18	71	469	64.58	367	20.15	18.36	17.30	0.11	0.01
< 30 Kb	70.05	72.80	783	65.73	499	23.94	21.79	20.60	0.16	0
< 40 Kb	71.97	74.97	1116	67.50	600	27.79	25.12	23.85	0.19	0
< 50 Kb	72.16	74.49	1364	67.54	687	30.46	27.52	26.28	0.19	0
< 60 Kb	72.98	74.83	1589	69.10	754	32.97	30.08	28.69	0.20	0
< 70 Kb	73.72	75.53	1782	809	809	35.13	32.09	30.45	0.20	0
< 80 Kb	74.24	75.95	1929	70.37	847	36.92	33.72	32.07	0.20	0
< 90 Kb	74.53	76.21	2068	70.60	881	38.43	35.06	33.57	0.20	0
< 1Mb	74.85	76.39	2181	71.14	901	39.46	36.18	34.70	0.21	0

Table 5 : taux d'usage des hotspots par résolution pour duohmm

Trois taux d'usage sont présentés : le taux d'usage global, paternel et maternel. Le nombre de crossing over pour les pères et les mères a été ajouté. Les colonnes chance sont les taux d'usage trouvés par hasard pour des perturbations de 200, 500 et 1000 Kb. Les deux dernières colonnes rendent compte du test de Kolmogorov Smirnov avec le résultat alpha du test et la p-valeur associée.

Il était cependant important de se renseigner sur la part des la variabilité des TML due à une origine génétique (autrement dit l'héritabilité du TML). L'héritabilité au caractère estimée par Gemma est de 3.19e-06 avec un écart-type de 1.024. En faisant fonctionner Gemma selon un modèle mixte bayésien on retrouve la même héritabilité estimée avec le même écart type. En revanche la distribution de l'héritabilité est plus informative **(Figure 16).** Au moins 70% de la distribution se situe au dessus de 0.25. Pour cette analyse ci, nous ne pouvons pas affirmer un chiffre précis pour l'héritabilité en revanche nous pouvons estimer qu'il se situe probablement entre 0.25 et 1. L'estimation la plus basse de 0.25 indique déjà une composante génétique très importante dans la variabilité du taux d'usage des points chauds.

2. *Analyse du taux d'usage des points chauds avec DuoHMM*

a) *taux d'usage estimé naïf*

L'analyse avec DuoHMM s'est faite dans un deuxième temps. Pour chaque crossing over supposé DuoHMM associe une probabilité de recombinaison **(annexe 10).** Nous avons choisi de ne prendre que les crossing over avec une probabilité supérieure à 0.9. En effet, avec cette probabilité la majorité des crossing over détectés par LinkPHASE sont retrouvés par DuoHMM **(Figure 17)**. L'approche DuoHMM a permis d'identifier 4831 crossing overs dans 65 individus (37 mâles et 28 femelles) et 350 méioses. Il y a moins de crossing over détectés par DuoHMM que par LinkPHASE en général. En revanche la résolution de DuoHMM est meilleure **(Figure 18).** Pour les intervals inférieurs à 30 Kb il y a 1282 crossing over (équivalent au nombre de crossing over détectés par LinkPHASE).

Le calcul du TN nous indique que pour des intervals inférieurs à 30 Kb le taux d'usage des points chauds est de 70.05% **(table 5)**. De plus le taux attendu par hasard est de seulement 23.94 % (moyenne des permutations de la position des crossing over **(annexe)**). Il

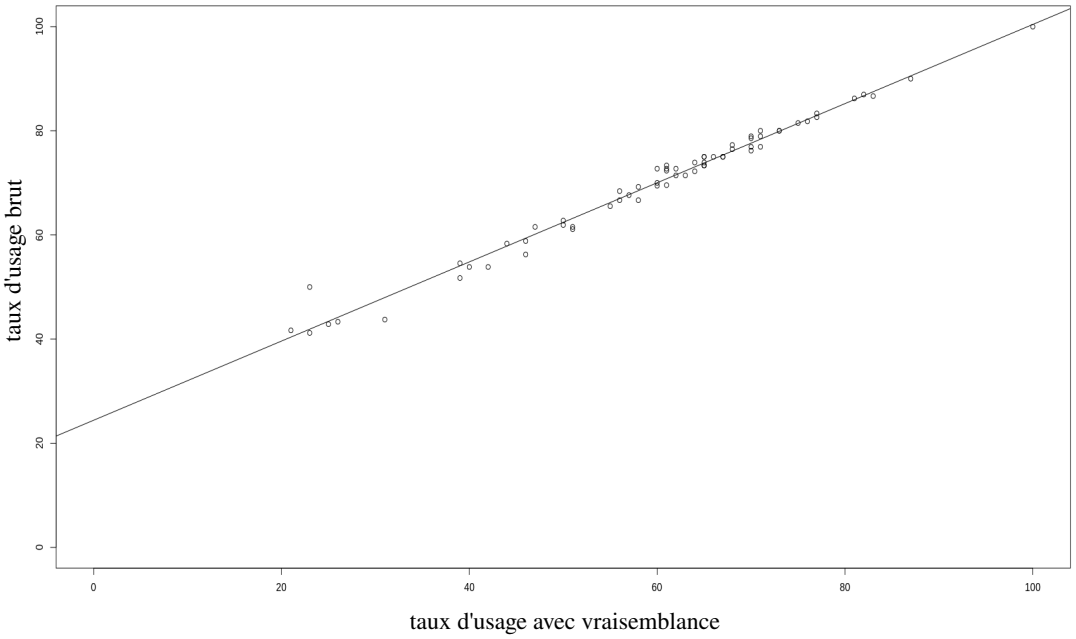


Figure 19 : taux d'usage corrigé alpha en fonction du taux d'usage brut pour Duohmm
Le taux d'usage corrigé est en en abscisse et le brut en ordonnée. Le nuage de point suit une droite qui nous indique que le taux brut surestime le taux corrigé surtout pour les valeurs faibles.

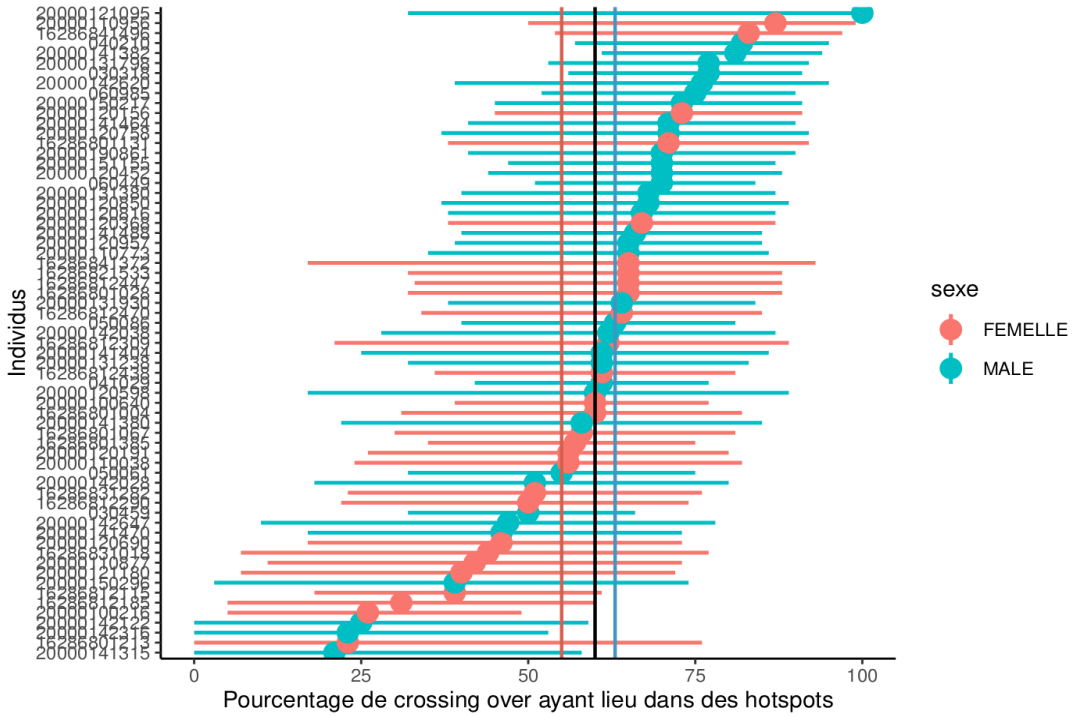


Figure 20 : pourcentage de cross-over apparaissant dans des hotspots par individu pour duohmm
Pourcentage de cross-overs détectés dans des intervalles de 30 kb qui ont lieu dans des hotspots basés sur une inférence de déséquilibre de liaison. Ces pourcentages sont calculés pour chaque individu dont les identifiants sont en ordonnée. L'estimation du maximum de vraisemblance est représentée par les points (bleus pour les mâles, rouges pour les femelles) tandis que les barres horizontales indiquent les intervalles de confiance (95 %). Les lignes verticales représentent l'estimation du maximum de vraisemblance pour l'ensemble des individus (noir), l'ensemble des femelles (rouge) et l'ensemble des mâles (bleu).

h0	LRT	ddl	p-value	pchisq
α_unique = α_sexes	5.26	1	0.023	0.0218
α_unique_mâle = α_indiv_mâles	40.34	36	0.395	0.28
α_unique_femelle = α_indiv_femelles	26.89	27	0.57	0.47
α_sexes = α_indiv_sexes	67.23	63	0.47	0.33

Table 6 : likelihood ratio pour les hypothèses sexes pour Duohmm
La première colonne fait état des différentes hypothèses nulles : est-ce qu'un taux d'usage unique est égal aux taux d'usage pour les mâles et les femelles? Est-ce que le taux d'usage d'un sexe est égale aux taux d'usage des individus ce sexe ? Enfin la somme des deux dernières hypothèses. L'avant dernière colonne est la p-valeur du likelihood ratio calculé avec 500 permutations d'intervalles. La dernière colonne est la p-valeur de la distribution du khi2 de ces permutations.

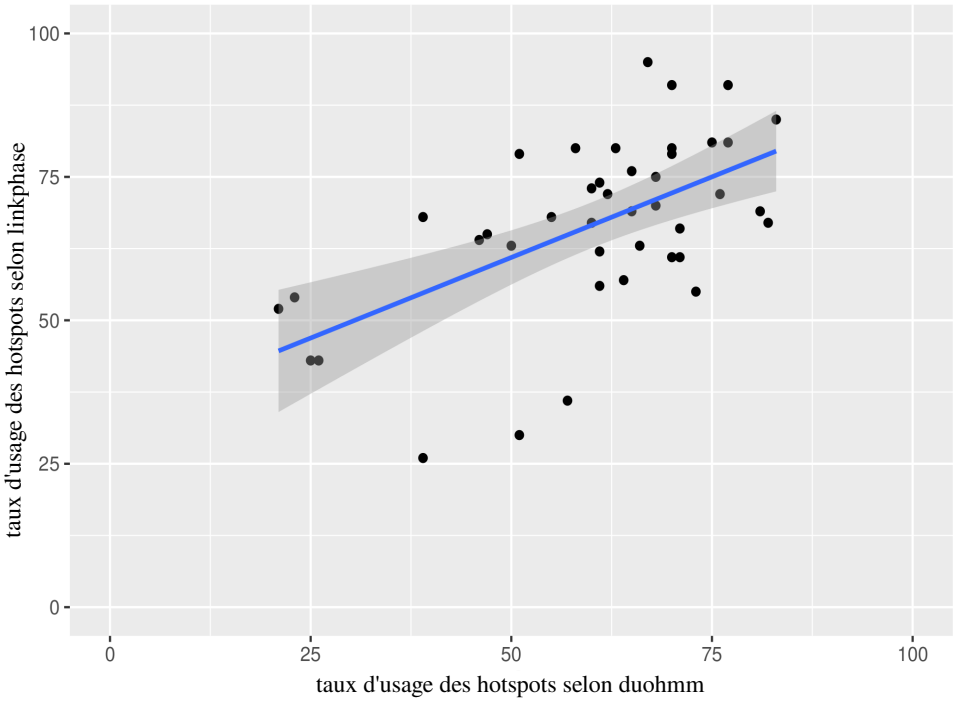
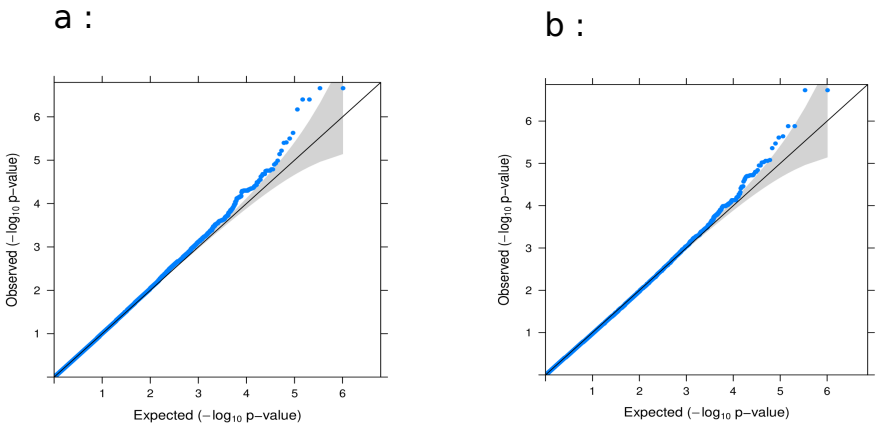


Figure 21 : taux d'usages corrigés des cross over detectés par duohmm en fonction de ceux detectés par Linkphase
En abscisse : taux d'usage corrigés par vraissemblance pour les données duohmm.
En ordonnée : taux d'usage corrigés par vraissemblance pour les données Linkphases.



a : modèle de base

b : modèle avec correction du sexe

Figure 22 : ajustement des distributions des p-valeurs des LRT à un modèle théorique pour Duohmm
Il s'agit des p-valeur du test LRT pour chaque SNP calculées par Gemma. En abscisse il s'agit des p-valeurs théoriques et en ordonnée les p-valeurs observées. La droite est la droite d'ajustement avec l'écart autorisé en gris. Pour avoir des données pertinentes il faut que le maximum de points suive la droite ou soit dans l'écart autorisé. Si quelques valeurs maximales se détachent de cette zone elles sont considérées significatives. 36 mâles et 28 femelles ont été utilisés.

y a encore une différence entre le taux d'usage des mâles et des femelles **(table 5)**. Le nombre de crossing over détectés est plus rapproché pour ces deux catégories qu'ils ne l'étaient dans LinkPHASE. En revanche le test de Kolmogorov-Smirnov ne nous permet pas de rejeter l'hypothèse selon laquelle distribution de la taille des intervalles est différente (ceux inférieurs à 30 Kb) **(table 5).**

b) taux d'usage estimé par maximum de vraisemblance

Le TN à tendance proposer des valeurs plus fortes que le TML de la même façon pour DuoHMM et LinkPHASE **(Figure 10 & Figure 19)**. Le TML varie entre individus (entre 25 et 100 % ; CI : 12 - 100 %) **(Figure 20)**. Il est de 60 % en moyenne. Il y a encore une différence entre mâles (taux : 63 %) et femelles (taux : 55 %). La comparaison entre les taux d'usages corrigés pour Linkphase et pour DuoHMM ne sont pas tout à fait identiques **(Figure 21).**

Pour confirmer ces résultats nous à nouveau avons calculé le rapport de vraisemblance pour plusieurs hypothèses et permuté 500 fois les crossing over entre individus **(table 6)**. L'hypothèse selon laquelle il y a une différence d'usage selon le sexe est significative (*p=0.023*). La différence d'usage des individus au sein des femelles est non significative (*p=0.56)* ainsi que chez les mâles (*p=0.40*). Les distributions des rapports des permutations pour les différentes hypothèse suivent encore une fois une loi du khi2 **(annexe.13)**.

c) GWAS

Nous avons réalisé une analyse d'association sur les 65 individus et les 27 chromosomes (dont le X). L'analyse d'association révèle que les p-valeurs sont calibrées pour le modèle de base (sans correction) et la correction du sexe **(Figure 22 : a, b)**. La correction avec le sexe semble améliorer un petit peu le modèle de base. De plus nous avons constaté

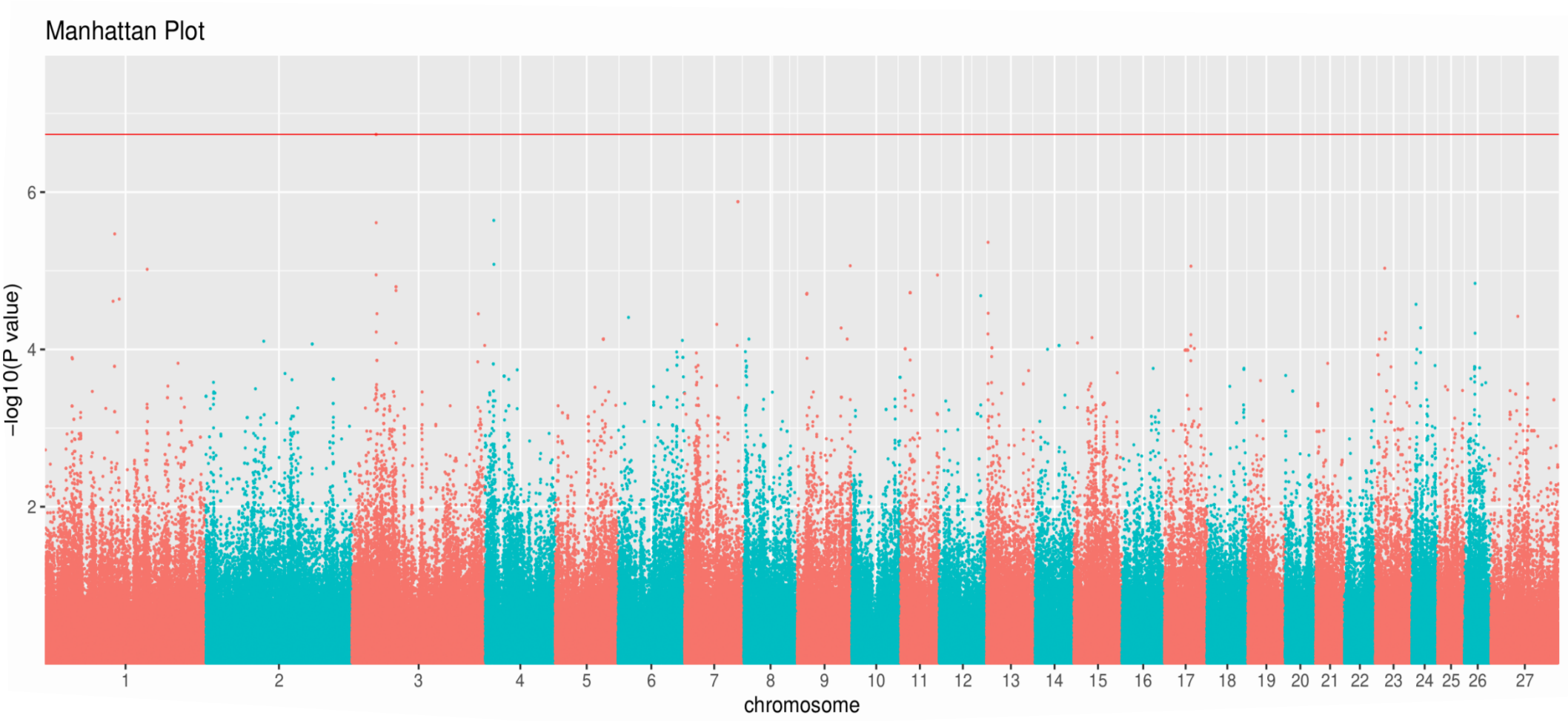
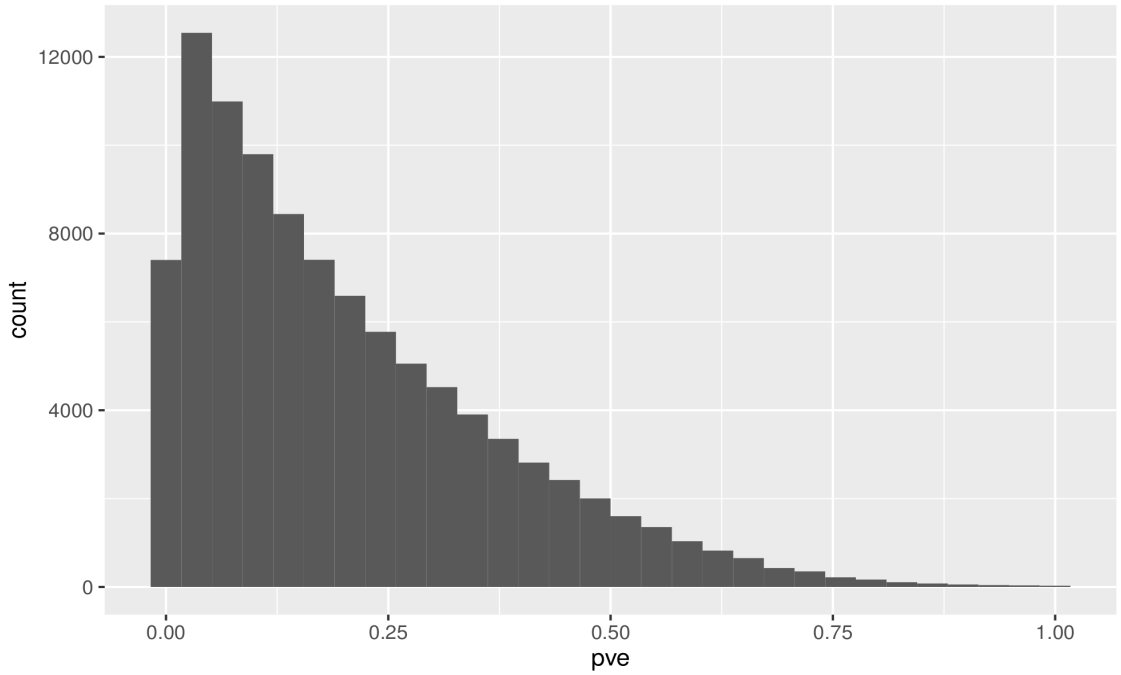


Figure 23 : manhattan plot des p-valeurs des effets des SNP pour Duohmm
En abscisse : les 26 autosomes et le chromosome X (27)
En ordonnées : les log10 des p-valeurs LRT des effets des SNP calculés par Gemma
Une valeur seuil a été placé avec un FDR de 0.05. 37 mâles et 28 femelles ont été utilisés et un effet sexe a été ajouté.

a :



b :

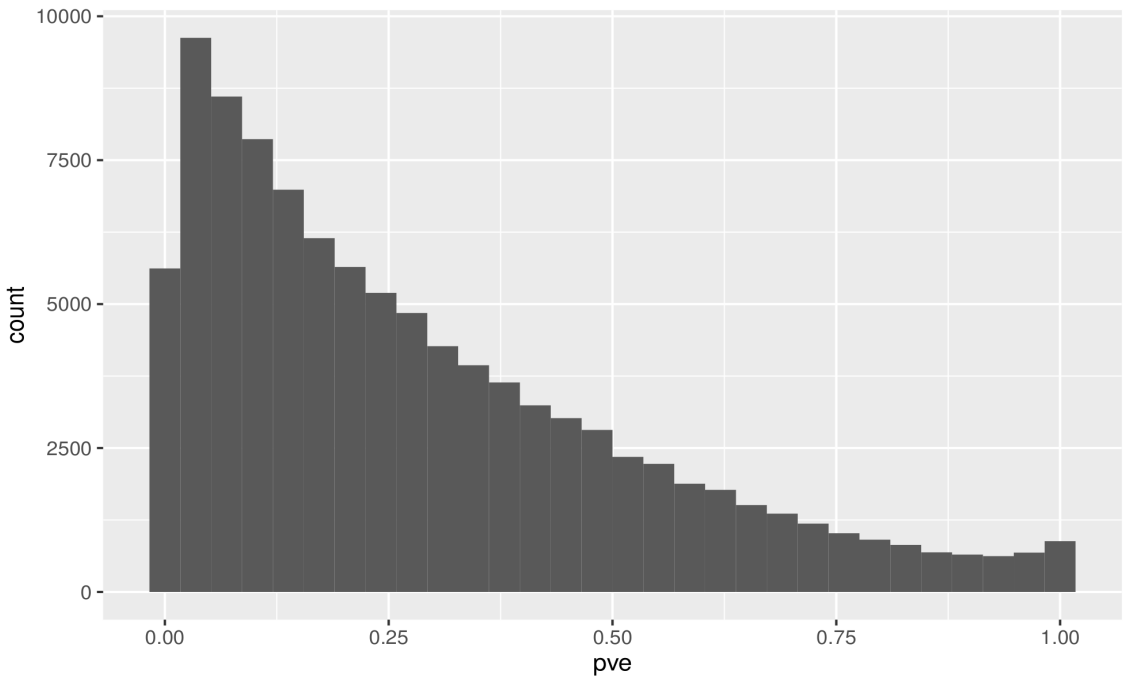


Figure 24 : distribution de l'héritabilité pour Duohmm
Distribution de l'héritabilité trouvée le modèle bayésien mixte de Gemma. En abscisse : pve ou estimation de l'héritabilité
Avec la fonction ecdf on trouve que plus de 70 % des héritabilités se situent au dessous de 0.25 (a) et 0.3 (b).

a : modèle avec un effet sexe sur les 65 individus
b : modèle classique avec 37 mâles

précédemment un effet sexe et nous avons suffisant de femelles contrairement à l'analyse LinkPHASE. C'est pourquoi nous allons nous concentrer sur le modèle avec une correction du sexe. Pour ce modèle deux SNP sont justes significatifs sur le chromosome 3 sur les positions 38401373 et 38401481 pour un FDR de 0.05 **(Figure 23)**. Nous avons regardé 1 Mb autours de ces positions sur UCSC Genome Browser (référence utilisée : Oar_v3.1/oviAri3). Les résultats montrent essentiellement des gènes de ménage **(annexe 8)**. L'un eux GMCL1 pourrait avoir un rôle dans la spermatogenèse durant le pachytène. C'est durant ce stade qu'ont lieu les crossing over. En revanche il semblerait plus être lié à la formation de l'enveloppe nucléaire. L'héritabilité n'est également pas informative (pve = 3.19e-06 , se = 1.024). Contrairement à LinkPHASE la distribution de l'héritabilité selon le modèle mixte bayésien nous indiquerait que l'héritabilité serait plus proche de 0 **(Figure 24)**. Pour l'ensemble des individus 70% de la distribution se situe en dessous de 0.25 **(Figure 24 : a)**. Pour les mâles 70% de la distribution se situe en dessous de 0.3 **(Figure 24 : b)**.

3) DISCUSSION

Le premier objectif de nos analyses était d'identifier si les crossing overs détectés dans les familles chevauchaient significativement les points chauds identifiés dans une étude précédente dans la populations Lacaune. Les résultats montrent sans ambiguïté que c'est bien le cas. Nous estimé les taux de recombinaisons en détectant les crossing over grâce à deux outils : LinkPHASE et DuoHMM. Pour ces deux outils le nombre de crossing over détectés avec une résolution de 30 Kb est comparable (LinkPHASE : 11095 ; DuoHMM : 4831). Nous avons réussi à caractériser les taux d'usage avec une assez bonne précision : la taille des intervalles comprenant les crossing over étudiés est de maximum 30 Kb et la caractérisation des points chauds a été faite avec une puce 600 K. Cette précision est comparable à celle des travaux de Coop et collaborateurs (2008) sur l'humain. D'ailleurs les

taux d'usages naïfs (TN) chez les mâles avec DuoHMM (72%) et LinkPHASE (78%) sont très proches de ceux de Coop et collaborateurs (75%) pour 30 Kb. Le taux d'usage global estimé par maximum de vraisemblance (TML) est de 60% pour DuoHMM ainsi que chez coop et al (2008). Il de 69 % pour LinkPHASE. L'estimation par maximum de vraisemblance semble donc avoir le même effet que chez l'humain : diminuer l'estimation du taux d'usage. Cette estimation est probablement plus conservative qu'une approche naïve car elle tient compte de la taille des crossing over ainsi que de la concentration en points chauds autour du crossing over. Nous avons pu appliquer cette méthode chez une autre espèce que l'homme avec succès.

Le deuxième objectif de cette étude était de savoir si des différences entre individus pour le taux d'usage des points chauds existent dans la population Romane. Pour LinkPHASE et DuoHMM le TML nous donne des variations entre 25 et 100%. Ces variations sont moins importantes que celles trouvées chez l'humain (entre 0 et 100%). Cependant elles ne sont pas significatives selon les permutations des crossing over entre individus. Les TML présentent des variations individuelles suggestives pour les femelles avec l'analyse LinkPHASE. Cependant il se trouve que cette analyse comprend 21 femelles de moins que l'analyse réalisée avec DuoHMM. Un autre élément à prendre en compte pour ces analyses est la différence de caractérisation des crossing over entre les deux outils. DuoHMM utilise une carte génétique comme information à priori contrairement à LinkPHASE. La conséquence en est une localisation plus fine (meilleure résolution). En revanche il est possible que DuoHMM homogénéise la localisation des crossing over en étant biaisé par les points chauds présents sur la carte. Ceci peut amener à un effacement artificiel des différences individuelles.

Si nous n'avons pas trouvé de différences individuelles il y a aurait en revanche un potentiel effet sexe : LinkPHASE (mâles : TML = 70 % et femelles : TML = 47% ; *p=0.0012*) et DuoHMM (mâles : TML = 63 % et femelles : TML = 55 % ; *p=0.023*). Ces résultats n'existaient pas chez Coop et collaborateurs (2008) qui n'avaient noté aucune différence entre hommes et femmes pour l'usage des points chauds. Dans une étude de 2018 (Brick et al. 2018) la différence de sexes dans l'usage des points chauds a été mise en évidence chez des souris. Les deux sexes utiliseraient les mêmes points chauds même certains seraient préférentiels aux mâles ou aux femelles. Cela serait dû aux méthylations (H3K4me3) présentes aux points chauds. Les auteurs expliquent également que la voie par défaut (se basant sur les méthylations), en absence de PRDM9 est plus utilisée chez les femelles.

Les écueils sur lesquels nous nous heurtons pour la différence des sexes sont d'une part les différences entre outils expliquées précédemment, d'autre part la descendance des moutons. En effet chez les animaux d'élevage les mâles ont plus de descendants. Dans cette étude les femelles ont entre 2 et 3 descendants tandis que les mâles en ont entre 2 et 29. De plus Coop et collaborateurs n'avaient que des familles nucléaires (les deux parents qui avaient donc le même nombre d'enfants) et un jeu de données plus important. Nous pourrions répéter nos analyses en conservant uniquement les familles nucléaires. Mais cela réduirait drastiquement le jeu de données ce qui pénaliserait la puissance de nos tests encore une fois.

Les derniers objectifs consistaient à explorer la nature génétique des différences d'usage des hotspots. Puisque les seules différences potentielles étaient de nature sexuelle nous nous sommes concentrés dessus. Nous n'avons pas pu associer le taux d'usage des points chauds au gène PRDM9 ou à un autre gène lié à la recombinaison. De fait nous avons trouvé des SNP significatifs sur le chromosome 3 (avec les données DuoHMM) mais aucun gène candidat fonctionnel pour le processus de recombinaison ne se trouvait aux alentours (1

Mb) de ces positions. Pour l'analyse LinkPHASE aucun SNP ne s'est révélé significatif. L'analyse d'association réalisée par Gemma sur LinkPHASE et DuoHMM a estimé une héritabilité de 3.19e-06 avec un écart type de 1.024. En utilisant les crossing over de LinkPHASE, la distribution de l'héritabilité du taux d'usage des points chauds indique une forte probabilité qu'elle soit comprise entre 0.25 et 1 (en utilisant le modèle mixte bayésien de Gemma). L'héritabilité trouvée avec DuoHMM tendrait plutôt à montrer l'inverse. Chez l'humain elle était de 0.22 (Coop et al. 2008) et chez le bovin de 0.21 (Sandor et al. 2012). L'homogénéisation de la localisation des crossing over par DuoHMM mentionnée plus haut peut biaiser notre estimation de l'héritabilité.

Pour l'instant ces résultats tendent à montrer que l'usage des points chauds n'est pas héritable chez le mouton et qu'il ne serait donc pas lié à PRDM9. Pour aller plus loin il faudrait affiner notre résolution des crossing over pour affiner notre phénotype. Le séquençage pourrait être une option. De plus un point non négligeable de notre travail est la taille du jeu de données. Le pari que nous avons pris est que un gène seul à effet majeur pourrait être détecté par une étude d'association même avec un faible effectif. Pour confirmer nos résultats il faudrait un jeu de données plus important. Si, dans de prochaines analyses, l'absence d'héritabilité et du rôle de PRDM9 se confirme il faudrait regarder où se situent les recombinaisons. Ont elles un phénotype du genre oiseaux? Ou un phénotype traditionnellement associé aux mammifères?

Au final nous avons caractérisé avec précision le taux d'usage des points chauds chez le mouton. Ce qui avait été fait uniquement chez l'humain. A présent il serait important d'approfondir nos résultats. La présence d'une variabilité inter-individuelle du taux d'usage des points-chauds dans la populations Romane, au delà d'un effet sexe potentiel, ne semble pas avérée d'après nos résultats. Cependant les différences d'estimation obtenues selon que

les crossing-overs ont été détectés avec LinkPHASE ou DuoHMM indique que cette étape de l'analyse est cruciale. Il conviendrait d'explorer plus avant les différences entre les deux logiciels pour identifier un jeu de crossing-overs robuste sur lequel faire une analyse intégrée. Par ailleurs, nous avons utilisé les points chauds détectés dans la race Lacaune. Il conviendrait de répliquer l'analyse avec des points chauds identifiés dans les races d'origine du croisement (Berrichon du Cher et Romanov).

Il pourrait être intéressant également de recommencer ces analyses sur un autre jeu de données plus important et/ou avec avec des familles nucléaires. Ainsi nos analyses révéleraient des différences plus précises de taux d'usages et nos analyses d'associations seraient plus efficaces.

Un autre aspect à considérer est le chromosome X. Il serait intéressant de pousser les analyses sur ce chromosome surtout si les différences de taux d'usages selon le sexe se vérifient. De plus le chromosome X possède une partie pseudo-autosomale. Or nous savons que chez la souris les mécanismes de recombinaison sont très différents dans cette zone (Baudat et al. 2013), (Brick et al. 2012). Il faudrait essayer d'analyser ces zones (pseudo et non pseudo autosomales) séparément.

Il pourrait également être intéressant de regarder les motifs des doigts de zinc des PRDM9 de notre jeu de données et d'essayer de les comparer aux motifs des points chauds.

RÉFÉRENCES

Ahlawat S, Sharma P, Sharma R, Arora R, Verma NK, Brahma B, Mishra P, De S. 2016. Evidence of positive selection and concerted evolution in the rapidly evolving PRDM9 zinc finger domain in goats and sheep. *Anim. Genet.* 47:740–751.

Baudat, Imai, de Massy. 2013. Meiotic recombination in mammals: localization and regulation. *Nat. Rev. Genet.*

Brick K, Thibault-Sennett S, Smagulova F, Lam K-WG, Pu Y, Pratto F, Camerini-Otero RD, Petukhova GV. 2018. Extensive sex differences at the initiation of genetic recombination. *Nature* 561:338–342

Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. 2012. Genetic recombination is directed away from functional genomic elements in mice. *Nature* 485:642–645.

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. *Science* 319:1395–1398.

Coop, Przeworski. 2007. An evolutionary view of human recombination. *Nat. Rev. Genet.*

Delaneau O, Marchini J, Zagury J-F. 2012. A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9:179–181.

Druet T, Georges M. 2015. LINKPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors. *Bioinformatics* 31:1677–1679.

Grey C, Baudat F, de Massy B. 2018. PRDM9, a driver of the genetic map.Cohen PE, editor. *PLOS Genet.* 14:e1007479.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29:217–222.

Johnston SE, Bérénos C, Slate J, Pemberton JM. 2016. Conserved Genetic Architecture Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep (*Ovis aries*). *Genetics* 203:583–598.

Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, Bickhart DM, Cole JB, Null DJ, Liu GE, et al. 2015. Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis.Auton AJ, editor. *PLOS Genet.* 11:e1005387.

Muñoz-Fuentes V, Marcet-Ortega M, Alkorta-Aranburu G, Linde Forsberg C, Morrell JM, Manzano-Piedras E, Söderberg A, Daniel K, Villalba A, Toth A, et al. 2015. Strong Artificial Selection in Domestic Mammals Did Not Result in an Increased Recombination Rate. *Mol. Biol. Evol.* 32:510–523.

Myers S. 2005. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science* 310:321–324.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, et al. 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness.Gibson G, editor. *PLoS Genet.* 10:e1004234.

Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. Recombination initiation maps of individual human genomes. *Science* 346:1256442–1256442.

Paigen K, Petkov P. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat. Rev. Genet.* 11:221–233.

Peñalba JV, Wolf JBW. 2020. From molecules to populations: appreciating and estimating recombination rate variation. *Nat. Rev. Genet.* [Internet]. Available from: http://www.nature.com/articles/s41576-020-0240-1

Petit M, Astruc J-M, Sarry J, Drouilhet L, Fabre S, Moreno C, Servin B. 2017. Variation in Recombination Rate and Its Genetic Determinism in Sheep Populations.

Genetics:genetics.300123.2017.
Ponting CP. 2011. What are the genomic drivers of the rapid evolution of PRDM9? *Trends Genet.* 27:165–171.
Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M. 2012. Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle.Paigen K, editor. *PLoS Genet.* 8:e1002854.
Schield, asquesi, Perry, Adams, Nikolakis, Westfall, Orton, Meik, Mackessy, Castoe. 2020. Snake recombination landscapes are concentrated in functional regions despite PRDM9. *Mol. Biol. Evol.*
Stephens M. 2016. False discovery rates: a new deal. *Biostatistics*:kxw041.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* 100:9440–9445.
Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44:821–824.