Méthodes statistiques pour la dissection de la variabilité des caractères à l'aide de puces SNP

MÉMOIRE

présenté et soutenu le 1 septembre 2011 pour l'obtention du diplôme de
MASTER 2 INGÉNIERIE MATHÉMATIQUES À TOULOUSE

par

Julien JACQUIN

Maîtres de stage : M. Hélène GILBERT, M. Jean-Michel ELSEN
Tuteur de stage : M. Jean-Marc AZAÏS

Méthodes statistiques pour la dissection de la variabilité des caractères à l'aide de puces SNP

MÉMOIRE

Julien JACQUIN
1 septembre 2011

Remerciements

Je remercie mes maîtres de stage Jean-Michel Elsen et Hélène Gilbert pour leurs précieux conseils et encadrement durant ce stage. Merci à M. Jean-Marc Azaïs pour ses conseils, qui m'ont été très utiles, à chaque fois que l'on s'est rencontré. Je remercie aussi Andrès Legarra, Simon Teyssedre et toute l'équipe méthode de la SAGA pour leurs divers avis et collaboration sur ce travail. Merci également à Rachel Rupp d'avoir partager son bureau avec ma personne pendant ces six derniers mois, merci à tous.

Table des matières

1 Introduction

1.1 Présentation de l'organisme d'acceuil

1.1.1 L'INRA

L'Institut National de la Recherche Agronomique, fondé en 1946, est le 1er institut de recherche agronomique européen avec près de 1800 chercheurs. C'est aussi l'un des trois premiers mondiaux en nombre de publications en sciences agricoles et en sciences des plantes et de l'animal. L'institut est placé sous le statut d'Etablissement Public à caractère Scientifique et Technologique(EPST) et sous la double tutelle du Ministère de la Recherche et du Ministère de l'Agriculture. L'INRA est constitué de 21 centres répartis en France métropolitaine et outre-mer.

1.1.2 La SAGA

Le lieu où s'est effectué mon stage est la Station d'Amélioration Génétique des Animaux (SAGA), créée en 1970 suite à la décentralisation du département de Génétique Animale et à la création du Centre de Toulouse. Sa mission est d'acquérir des connaissances afin de contribuer à l'amélioration génétique des ovins, des caprins, des lapins, des palmipèdes gras et des équins.
- L'activité de la SAGA s'organise autour de 3 axes :
- L'étude de la variabilité génétique des caractères d'intérêt zootechnique : la résistance aux maladies, la reproduction, la croissance et les aptitudes bouchères, la lactation et la qualité du lait, les phanères et le comportement et l'adaptation,
- Le développement des modèles, des méthodes et outils de la sélection animale,
- Les méthodes de gestion des populations animales,

1.2 Le stage

1.2.1 Contexte et objectifs du sujet de stage

Mon stage s'inscrit dans le cadre de la sous tâche 1 de la Tâche 2 du projet " Rules & Tools " financé par l'ANR (Association Nationale de la Recherche), qui vise à améliorer les méthodes de prise en compte du déséquilibre de liaison (DL) pour la cartographie de QTL, et l'estimation des probabilités d'être IBD entre segments chromosomiques. J'ai pu commencer mon stage vers mi-mars 2011 en étant très bien accueilli à la SAGA avec un ordinateur personnel à disposition. Durant toute la période de mon stage j'ai été encadré par Hélène Gilbert (GABI-LGC), chercheur confirmé (CR1), et Jean-Michel Elsen (SAGA), Directeur de recherche, qui m'ont beaucoup fait bénéficier de leurs expériences et savoirs respectifs dans le domaine de la génétique. J'ai été inséré dans l'équipe dite "méthode" qui vise au développement des modèles, des méthodes et des outils pour la sélection animale. Ainsi j'ai pu assister à des réunions méthodologiques de semaine en semaine qui m'ont permis de mieux cerner les outils de la sélection animale.

1.2.2 Les divers environnements et langages utilisés

Durant ce stage j'ai eu à programmer tous mes algorithmes en Fortran 90 en ayant une utilisation intensive de ce dernier. Ce choix vient du fait que la plupart des programmes utilisés en routine dans le domaine de la génétique animale sont écrits dans ce

langage. Très heureusement j'ai eu accès à un compilateur Fortran, par l'ouverture d'un compte Unix, sur l'une des plate-formes du GIS Genotoul ; la plate-forme bio-informatique (http ://bioinfo.genotoul.fr/). La puissance de calcul de cette plate-forme est apportée par différents serveurs et un cluster de 350 coeurs acquis en 2009. Pour les représentations graphiques et le prototypage de certains algorithmes j'ai utilisé MatlabR2009b en complément du Fortran90. Il m'est arrivé aussi, très rarement, d'utiliser le logiciel *R* pour quelques représentations graphiques.

1.2.3 Rencontre avec les membres du projet " Rules & Tools"

Au 31 mars j'ai eu l'opportunité de rencontrer les membres du projet " Rules & Tools", à INRA Jouy-en-Josas, lors de la troisième réunion du consortium. De ce fait j'ai fait la rencontre des différentes équipes travaillant sur une tâche respective du projet. Il existe 7 tâches (et sous-tâches associées) à ce projet et j'ai donc pu prendre connaissance des différentes relations, et implications, entre chacune de ces tâches lors de cette rencontre. En outre j'ai eu l'opportunité, au mois de juillet, de présenter l'état d'avancement de mes travaux lors d'une réunion dans le cadre de la tâche 2 de "Rules & Tools".

2 Vocabulaire élémentaire et notions de génétique mendélienne

2.1 Vocabulaire élémentaire

Marqueurs génétiques, locus et allèles

Un chromosome peut être modélisé par un intervalle réel. Un marqueur génétique est une séquence d'ADN spécifiquement repérable à laquelle on associe sa position (son locus) par rapport à l'origine (le début) du chromosome. Les marqueurs génétiques sont utiles afin de baliser un chromosome et d'en avoir une description. Il en existe divers mais les deux plus utilisés en génétique animale sont les SNP (*Single Nucleotide Polymorphism*) et les MST (*Microsatellite*). Ces marqueurs sont caractérisés par des variations dans une séquence de bases (exemple : TGTGTGTG, TATATATA) dans le cas des MST ou des variations pour une seule base (A, T, C ou G) dans le cas des SNP. On appelle allèle la variation spécifique du marqueur. Dans le cas des SNP il y a confusion entre une base et un allèle. Il est à noter que les SNP sont très souvent bialléliques, c'est à dire qu'un SNP peut être soit A1 ou A2 (exclusivement) à un marqueur. Exemple : A ou G pour un marqueur i, T ou A pour un marqueur i', avec $i \neq i'$.

2.2 Notions de génétique mendélienne

2.2.1 Le processus de méiose et la recombinaison

Les organismes diploïdes à reproduction sexuée sont caractérisés par n paires de chromosomes homologues, chaque élément de la paire étant reçu d'un gamète haploïde qui est une cellule sexuelle provenant de la méiose. La méiose quant à elle est un type de division cellulaire qui s'effectue chez les organismes à reproduction sexuée afin d'aboutir à la production des gamètes. Ces gamètes contiennent donc n chromosomes afin de restaurer les n paires de chromosomes durant la reproduction. Durant le processus de méiose il peut

y avoir un échange de matériel génétique entre les chromosomes homologues d'une paire. Cet échange est ce que l'on appelle le phénomène de recombinaison ou " Crossing-Over " (figure 1) et il peut arriver à plusieurs endroits entre les deux chromosomes.

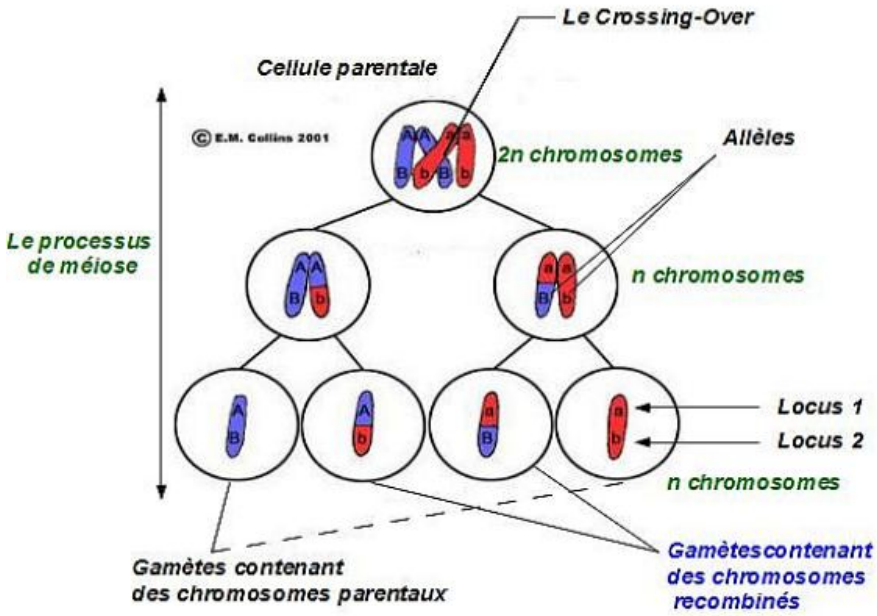

FIGURE 1 – **La recombinaison**

2.2.2 La notion de génotype, d'haplotype, de phénotype et de QTL

A chacun des marqueurs sur un chromosome, le chomosome paternel par exemple, il existe soit le même allèle à ce marqueur sur le deuxième chromosome homologue, le chromosome maternel dans ce cas, ou un autre allèle. On parle alors du génotype à ce marqueur pour la paire d'allèles reçu des parents. Pour plusieurs allèles voisins sur un chromosome (paternel ou maternel) on parle alors d'haplotype ou de phase. Par exemple les 4 couples d'allèles AB, Ab, aB et ab obtenus précédemment (figure 1) définissent des haplotypes. Associé au génotype à un locus il existe, parfois, un effet discriminant sur un caractère visible d'un organisme diploïde. La couleur des yeux chez les êtres humains en est un exemple. Le caractère visible exprimé qui est la couleur des yeux est appelé le phénotype de l'individu et est très différent selon le génotype de celui-ci. De même il existe des caractères quantitatifs d'intérêt chez les animaux d'élevage, tel que le poids et la taille, pour lesquels on voudrait localiser le locus ayant un effet sur ce dernier. Un tel locus est ce que l'on appelle un QTL pour " Quantitative Trait Locus ".

2.2.3 Le taux de recombinaison et la distance génétique

Le taux de recombinaison r entre 2 loci est défini comme la probabilité d'avoir une recombinaison entre ces 2 loci au cours d'une méiose. On a une plus faible probabilité de recombiner entre deux loci qui sont physiquement proches. Le taux de recombinaison fournit donc une bonne mesure de la distance entre loci. Toutefois cette distance n'est pas additive car une recombinaison déjà réalisée empêche l'occurrence d'une autre recom-

binaison à proximité (phénomène d'interférence). La distance génétique est plutôt donc définie comme une fonction du taux de recombinaison. La distance génétique entre 2 loci la plus couramment utilisée est la distance de Haldane (1919) :

$$d = \begin{cases} -\frac{1}{2}ln(1-2r) & si\ \ 0 \leq r < \frac{1}{2} \\ +\infty & sinon \end{cases}$$

Elle est obtenue en supposant que les recombinaisons sont indépendantes et que leur nombre suit un processus de Poisson d'intensité 1, même si dans la réalité ce n'est pas le cas à cause du phénomène d'interférence. L'unité de mesure de la distance génétique est le $centiMorgan$ (cM) et 1 cM correspond à un taux de recombinaison de $0,01$ entre 2 loci lors d'une méiose. Sachant que certains chromosomes chez les mammifères d'élevage (ou l'homme) ont une taille avoisinante à 100 cM on observe en espérance une recombinaison (au moins) à un endroit du chromosome après une méiose.

2.2.4 La notion de déséquilibre de liaison (DL) entre marqueurs génétiques

Le terme de déséquilibre de liaison (DL) entre marqueurs génétiques traduit l'association préférentielle qu'il peut y avoir entre les allèles de ces derniers. En d'autres termes il traduit l'indépendance probabiliste existante, ou non, entre les marqueurs. Si on considère 2 marqueurs génétiques biallélliques A et B ayant pour allèles (A_1, A_2) et (B_1, B_2) respectivement, alors on obtient un ensemble de 4 haplotypes possibles définis par ces 2 couples d'allèles aux 2 marqueurs : $\{A_1B_1, A_1B_2, A_2B_1, A_2B_2\}$. Les tableaux ci-dessous décrivent les relations entre les fréquences des haplotypes et celles de chacun des allèles :

Haplotypes	*Fréquences*
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

Tableau 1 : Les fréquences haplotypiques

Allèles	*Fréquences*
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

Tableau 2 : Les fréquences alléliques : estimation à partir des fréquences haplotypiques

Comme les fréquences sont relatives à la taille de la population on a $p_1 + p_2 = 1$ et $q_1 + q_2 = 1$, i.e : $\sum_{i,j} x_{ij} = 1$, par construction. Si les allèles aux deux marqueurs A et B s'associent aléatoirement alors les fréquences des 4 haplotypes (fréquences haplotypiques) seront égales aux produit des fréquences des allèles (fréquences alléliques) portés par ces haplotypes, sinon on aura une déviation, une quantité qu'on notera D, dans les fréquences haplotypiques.

Espérance des Fréquences
$x_{11} = p_1q_1$
$x_{12} = p_1q_2$
$x_{22} = p_2q_2$
$x_{21} = p_2q_1$

Tableau 3 : Espérances des fréquences sous l'hypothèse d'association aléatoire

Espérance des Fréquences
$x_{11} = p_1q_1 + D$
$x_{12} = p_1q_2 - D$
$x_{22} = p_2q_2 + D$
$x_{21} = p_2q_1 - D$

Tableau 4 : La déviation D dans le cas d'association non aléatoire

Cette quantité D est le déséquilibre de liaison. On vérifie aisément des expressions du *Tableau 4* que $D = x_{11}x_{22} - x_{12}x_{21}$. D traduit donc la différence fréquentiste qui existe entre les gamètes porteurs des haplotypes A_1B_1 et A_2B_2 de ceux portant les haplotypes A_1B_2 et A_2B_1, d'où le terme déséquilibre gamétique employé quelques fois dans la littérature. Ce coefficient est une quantité qui varie dans l'intervalle $[-0,25; 0,25]$, le maximum et le minimum de l'intervalle étant atteint pour $x_{11} = x_{22} = 0,5$ et $x_{12} = x_{21} = 0,5$ respectivement. En effet si $x_{11} = x_{22} = 0,5$, sachant que $\sum_{i,j} x_{ij} = 1$, alors $x_{12} = x_{21} = 0$ car $\forall(i,j)$ $x_{ij} \geq 0$, i.e : A_1B_2 et A_2B_1 ne peuvent pas exister dans la population et réciproquement pour le minimum. Comme les fréquences haplotypiques x_{ij} sont toujours positives il vient aussi que la valeur maximale et minimale de D dépendent des fréquences alléliques dans la population. En effet on a ;

$$\begin{cases} x_{12} = p_1q_2 - D \geq 0 & \text{i.e } D \leq p_1q_2 \ (1) \\ x_{21} = p_2q_1 - D \geq 0 & \text{i.e } D \leq p_2q_1 \ (2) \end{cases}$$

Comme les deux conditions (1) et (2) ne peuvent pas être vérifiées en même temps on a forcément que $D \leq min(p_1q_2, p_2q_1)$ pour la valeur maximale de D. Par symétrie on obtient aussi $D \geq max(-p_1q_1, -p_2q_2)$ pour la valeur minimale i.e $D \geq -min(p_1q_1, p_2q_2)$.

Une valeur normalisée D' du DL est donc obtenue par le rapport suivant (Lewontin,1964) :

$$D' = \frac{|D|}{D_{max}} \text{ tel que } D' \in [0; 1]$$

où $D_{max} = \begin{cases} min(p_1q_2, p_2q_1) & \text{si } D > 0 \\ min(p_1q_1, p_2q_2) & \text{sinon.} \end{cases}$

Il existe plusieurs mesures du DL entre marqueurs génétiques, comme celle du r^2 (Hill & Robertson, 1968) définie de la façon suivante :

$$r^2 = \frac{D^2}{p_1p_2q_1q_2}$$

C'est une mesure de référence pour les loci bialléliques, comprise entre 0 et 1, mais elle est très dépendante des fréquences allèliques et ne prend la valeur 1 que si chaque allèle du marqueur A est associé à un allèle unique au marqueur B. La mesure du D' quant à elle est celle qui est la moins dépendante des fréquences alléliques mais reste très dépendante des fréquences haplotypiques. Toutefois une étude de Heifetz et al. (2005), s'appuyant sur 3 lignées commerciales de poules pondeuses, met en évidence une probable surévaluation de l'étendue du DL lorsque celui-ci est évalué avec le D'. Ils ont également constaté que la réduction de la taille des échantillons d'individus augmentait le nombre de valeurs de D' très fortes, probablement à cause d'haplotypes manquants dans l'échantillon. Néanmoins la mesure qui reste la plus utilisée jusqu'ici en population animale, et que j'ai le plus utilisé lors de mon stage, est le D'.

2.2.5 Le DL, la recombinaison et les causes du DL

Soit $D_0 = x_{11} - p_1q_1$ le DL initial existant entre les deux marqueurs A et B dans une population animale et r le taux de recombinaison entre ces 2 marqueurs. Si on suppose qu'il n'existe pas de variation des fréquences alléliques entre 2 générations (le principe d'équilibre d'Hardy Weinberg, HWE) alors le DL diminue d'un facteur de $(1-r)$ à chaque génération pour des accouplements aléatoires.

En effet, soit x'_{11} la fréquence haplotypique de A_1B_1 à la génération suivante, on a $D_1 = x'_{11} = (1-r)x_{11} + rp_1q_1$ car $(1-r)$ est la probabilité de non recombinaison entre les allèles A_1 et B_1et rp_1q_1 représente la probabilité qu'on ait un haplotype A_1B_1 avec une recombinaison entre les allèles A_1 et B_1. Comme l'évènement d'avoir l'allèle A_1 (de probabilité p_1), celui d'avoir B_1 (de probabilité q_1) et celui d'avoir une recombinaison (de probabilité r) sont des évènements indépendants il suffit de faire le produit des 3 probabilités. Or $x'_{11} = (1-r)x_{11} + rp_1q_1 \Leftrightarrow x'_{11} - p_1q_1 = (1-r)x_{11} + p_1q_1(r-1)$, d'où $D_1 = (1-r)(x_{11} - p_1q_1) = (1-r)D_0$. Par principe de récurrence on a donc $D_t = (1-r)^t D_0$, où D_t est le DL calculé à la t-ième génération. On peut aussi approximer ce résultat par $D_t = e^{-rt}D_0$.
Il est à noter aussi que la distance génétique (d) est une fonction croissante du taux de recombinaison r et il vient donc que D_t décroît en fonction de cette distance.

Parmi les différentes causes du déséquilibre de liaison on peut citer :

- Le processus de mutation : la création d'un nouvel allèle à un locus crée un déséquilibre avec les allèles des marqueurs adjacents dans une population.
- La sélection animale, qui joue un rôle très important, car on choisit des reproducteurs qui transmettront leurs chromosomes successivement dans une population. Cela conduit donc à une réduction de la diversité génétique, et à un déséquilibre substantiel, à certains endroits du génome.
- La dérive génétique, qui est le processus par lequel les fréquences alléliques changent, certains allèles disparaissent, dans les populations de tailles finies à cause de biais aléatoires d'échantillonnage dans la transmission d'allèles d'une génération à l'autre.
- Le mélange de deux populations en équilibre de liaison qui constitue un ensemble globalement en déséquilibre de liaison dès lors que leurs fréquences alléliques sont respectivement différentes.

2.2.6 La notion d'identité par descendance (IBD) et par état (IBS)

On dit que deux segments (ou portions) chromosomiques sont IBD(“ Identical by Descent”) s'ils ont été hérités d'un même chromosome ancestral. En absence de mutation ils possèdent donc les mêmes allèles. Et deux segments chromosomiques sont dits IBS(“ Identical by State”) s'ils possèdent les mêmes allèles.
A un locus ou un marqueur on a donc localement :

IBD $\Rightarrow$ IBS
nonIBS $\Rightarrow$ nonIBD

En effet cette propriété n'est valable que localement car deux chromosomes nonIBS peuvent partager des segments IBD.

2.2.7 L'intérêt de ces notions en cartographie de QTL

L'idée importante à partir de ces notions est que la recherche d'un segment chromosomique IBD où la mutation causale (l'allèle ayant un effet favorable issu de la mutation) est apparue, chez des animaux présentant le même phénotype favorable, permettrait de localiser le QTL. Pour cela on suppose ***qu'après n générations de recombinaisons successives le DL est élevé dans une région petite IBD autour du QTL***, c'est l'hypothèse centrale qui justifie toute la démarche que j'ai eu durant mon stage. La figure 2 ci-dessous illustre cette théorie.

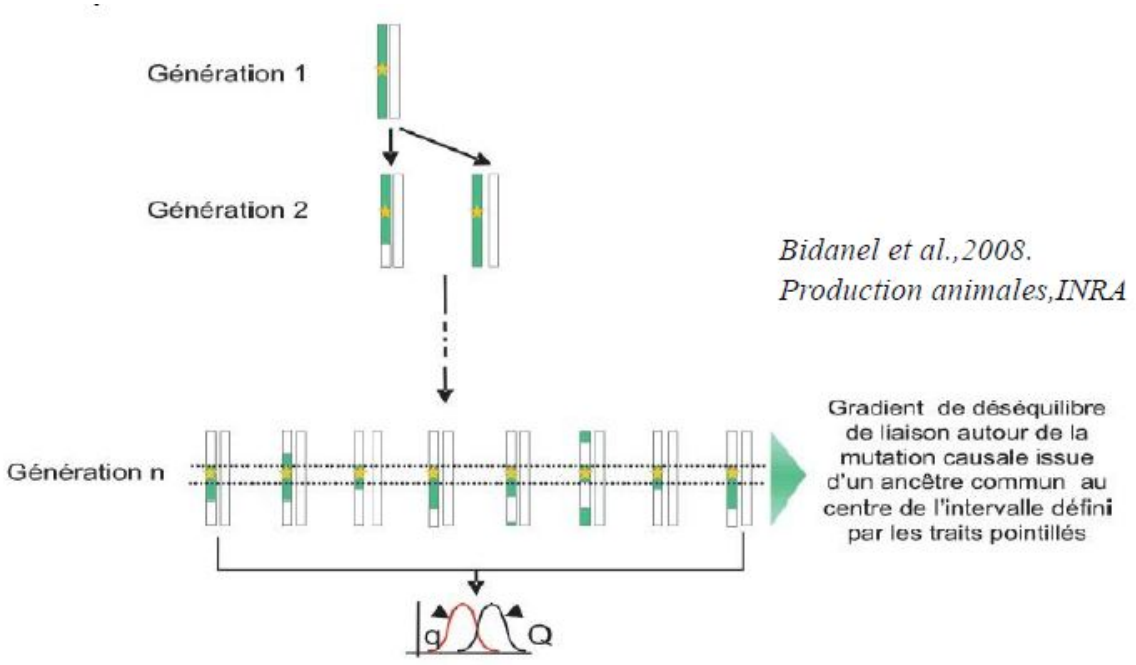

FIGURE 2 – **Le gradient du DL dans une région IBD petite**

Comme il est impossible de savoir si deux régions sont IBD il est nécessaire d'exploiter des modélisations de la ressemblance entre segments chromosomiques telle que l'IBS (la plus simple), le Score de Similarité (Li et Jiang, 2005), $\mathbb{P}$(IBD) ; la probabilité d'être IBD (Meuwissen & Goddard, 2001), et les HMM ("Hidden Markov Models"). On référa à s(IBD) pour les méthodes de similitude qui seront étudiés.

3 Matériels et méthodes

3.1 Le pedigree et les chromosomes de base pour le pedigree

3.1.1 La notion de pedigree et de "genedropping" à partir de fondateurs d'un pedigree

Un pedigree est un arbre où les individus aux sommets, dont on ne connaît pas les parents, sont appelés les fondateurs du pedigree. Une représentation possible d'un pedigree est donnée par le tableau suivant :

Ind	*Père*	*Mère*
1	0	0
2	0	0
3	1	2
4	1	3
5	3	4
⋮	⋮	⋮

Les individus 1 et 2 sont des fondateurs car leurs parents sont identifiés à 0 (parent inconnu). Un parent dans le pedigree est caractérisé par le fait qu'il se situe dans les colonnes Père ou Mère (3 ou 4 par exemple). A une génération un parent transmet la moitié de son matériel génétique (n chromosomes) à un individu dans le pedigree. Si on est capable d'observer et de faire la distinction entre les chromosomes des fondateurs alors on sera capable de connaître l'origine des segments chromosomiques (formés par le processus de méiose) reçus par les descendants. Ce processus de transmission simulée de segments chromosomiques formés par des recombinaisons sucessives au cours des générations est appelé "genedropping". La figure 3 ci-dessous illustre ce concept :

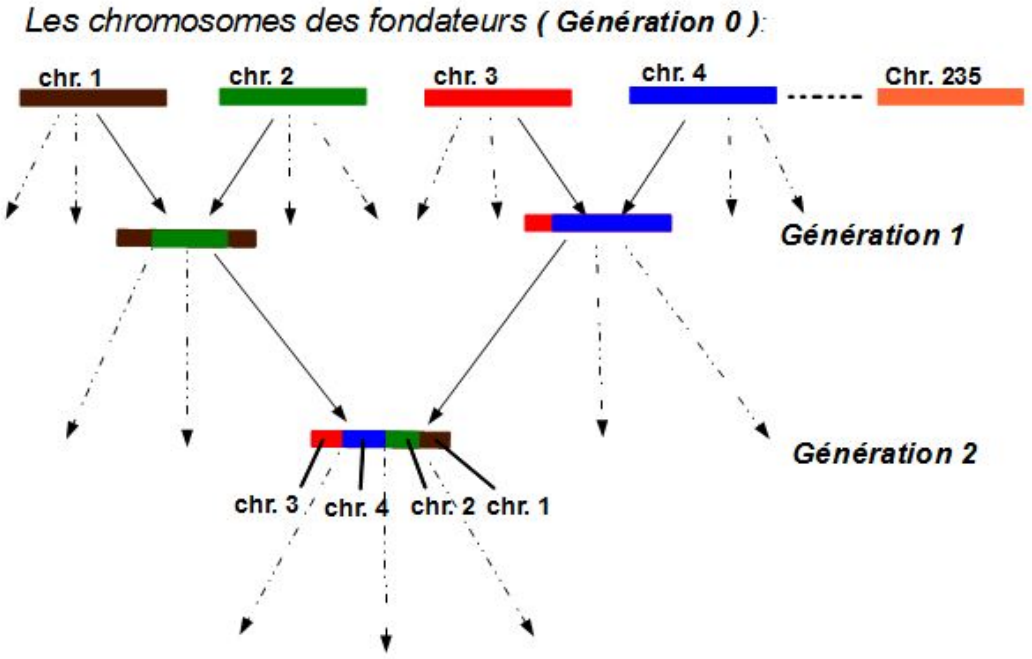

FIGURE 3 – **Le processus de "genedropping"**

Le pedigree et l'ensemble de chromosomes de base qu'on utilise pour les simulations sont des données réelles. L'ensemble de chromosomes réels sert de pool de gamètes à affecter aux individus fondateurs du pedigree. Le simulateur qui sert à affecter ces chromosomes aux fondateurs et à effectuer le "genedropping" se nomme *Pedigree*. L'utilisation de données réelles a pour objectif de produire des simulations aussi réalistes que possible pour ensuite comparer les mesures de similitude entre certaines régions chromosomiques à des statuts IBD entre ces régions.

3.1.2 Le pedigree

Le pedigree est constitué de 1594 fondateurs, 3373 pères, 7100 mères et 1282 descendants (non reproducteurs), animaux nés entre 1970 et 2004 dans la population de porcs de la race Large White en France. Dans les 25 premières générations, seuls les reproducteurs ont été conservés. La génération 26 est constituée des 1282 descendants. ***Dans ce pedigree, pour se rapprocher d'une situation réelle, on considère $m = 485$ individus entre 1996 et 2004 comme étant les fondateurs pour la cartographie de QTL, ce sont les parents d'animaux nés entre 2000 et 2004.*** En pratique on connaît les génotypes des derniers individus nés dans un pedigree, ici on a supposé qu'on ne connaissait pas les génotypes des individus pour la période de 1970 à 1996. La figure 4 ci-dessous illustre le pedigree utilisé pour les simulations.

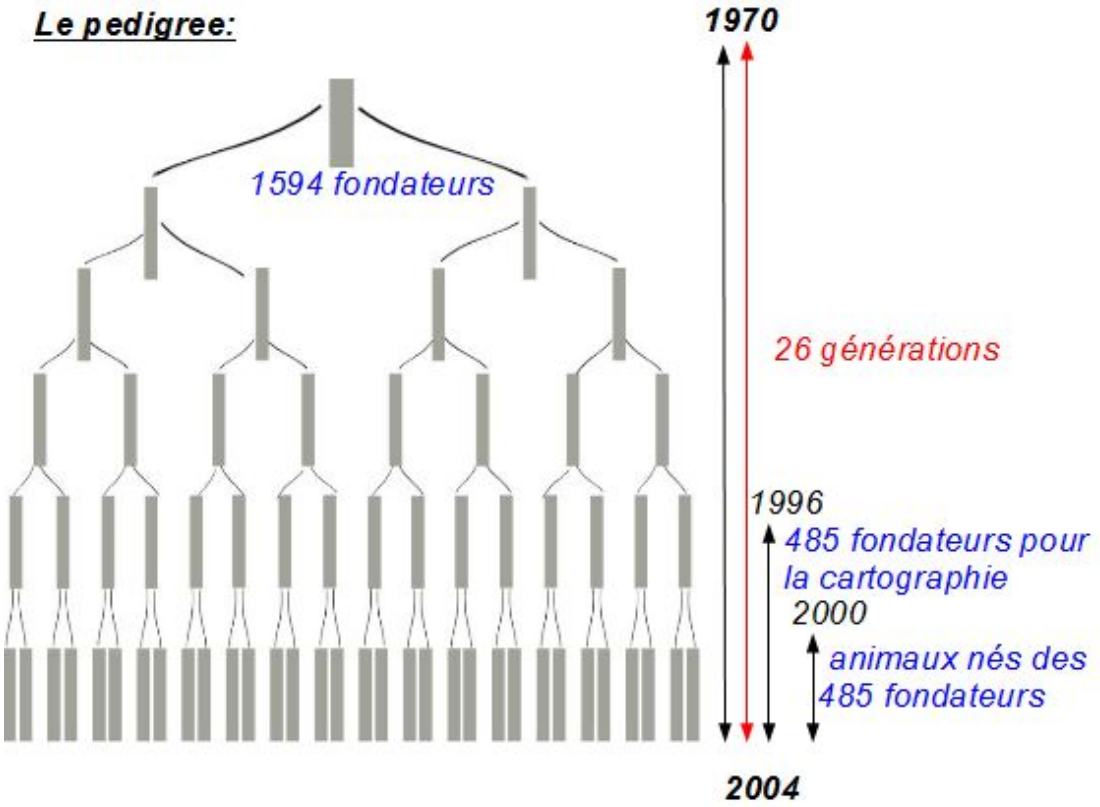

FIGURE 4 – **Le pedigree**

3.1.3 L'ensemble de chromosomes de base

Les chromosomes de base est un ensemble de ***235 chromosomes*** *(SSC18 : Chromosome 18 porcin)* construits dans le cadre du projet DELISUS par *Bertrand Servin (LGC, INRA)*. De cet ensemble on puise des chromosomes par des tirages aléatoires uniforme avec remise pour les affecter aux 1594 fondateurs du pedigree. ***Les chromosomes sont balisés avec 1252 SNP et après suppression des marqueurs présentant une "Minor Allele Frequency" (MAF) inférieure à*** 5% ***on ne garde que 969 SNP.*** La MAF est la fréquence de l'allèle le moins représenté (à un marqueur) dans la population. Ces fréquences faibles affectent le calcul du DL avec les autres marqueurs, il est donc important de ne pas les considérer pour le calcul du DL.

3.1.4 Le traçage des segments chromosomiques selon leurs origines

Afin de connaître l'origine chromosomique des 969 SNP d'un chromosome à une génération quelconque les 969 SNP sont tous numérotés (procédé appelé "tagging") de 1 à 235 selon leur chromosome réel d'origine. Les processus de transmission à chaque génération depuis les 1594 fondateurs du pedigree sont donc accompagnés d'un processus dual de transmission qui tient juste compte de l'origine chromosomique du marqueur (SNP), c'est le processus de "genedropping" cité précédemment.
La figure 5 ci-dessous illustre la notion d'origine chromosomique des allèles pour les 6 premiers marqueurs, sur chacun des chromosomes l et l', pour un individu à la génération n. Les allèles aux marqueurs M_1 et M_2 ont été reçus du chromosome réel 58 et 230 respectivement pour le chromosome $l^{'}$ par exemple. Il est à noter aussi que les 2 allèles à chacun des marqueurs ont été codés en 1 et 2 avec une table de correspondance qui donne la base correspondant à 1 et à 2, par exemple au marqueur M_3 on a 1 $\Leftrightarrow$"T" et 2 $\Leftrightarrow$ "C".

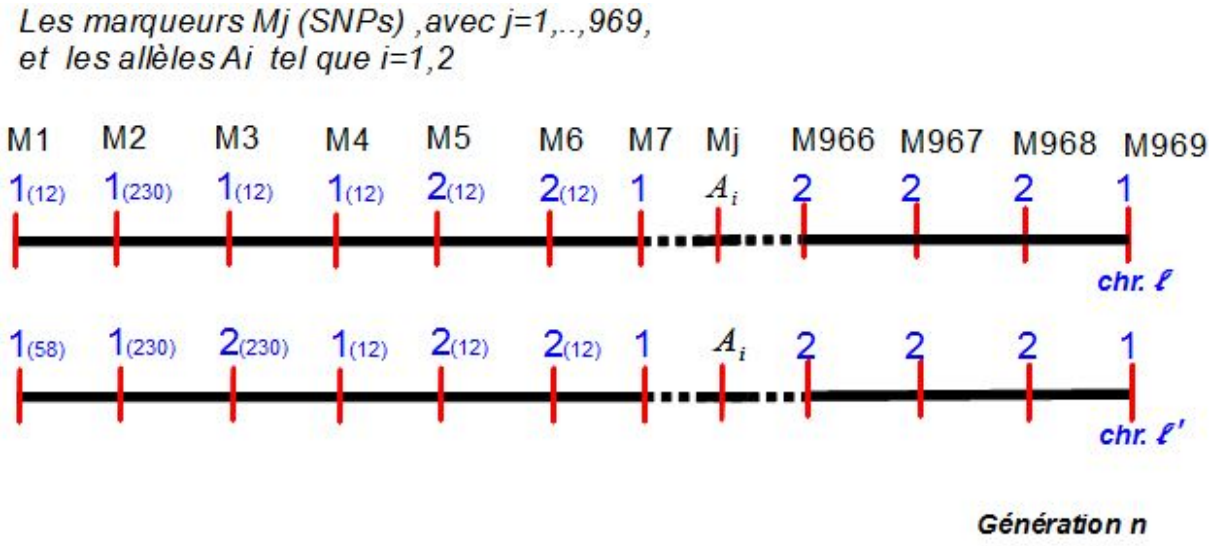

FIGURE 5 – **Les 969 SNPs biallélіques**

3.1.5 La description du DL pour l'ensemble de chromosomes de base

Comme il a été précisé au début de ce mémoire le $D^{'}$ de Lewontin (1964) est la mesure du DL qui est la plus utilisée jusqu'ici en population animale. Le calcul du DL par la mesure du $D^{'}$ pour les 235 chromosomes de base est représenté par le graphique (figure 6) ci-dessous. Le graphique représente le gradient du DL calculé entre les 969 SNP deux à deux pour ces 235 chromosomes. On remarque que le DL est plus prononcé à certains endroits qu'à d'autres. Une description comme celle de la figure 7, parmi d'autres possibles, permet de regarder les profils de DL à gauche et à droite autour de certains marqueurs qu'on choisit pour simuler la présence d'un QTL afin d'évaluer les méthodes s(IBD) utilisés pour la cartographie.

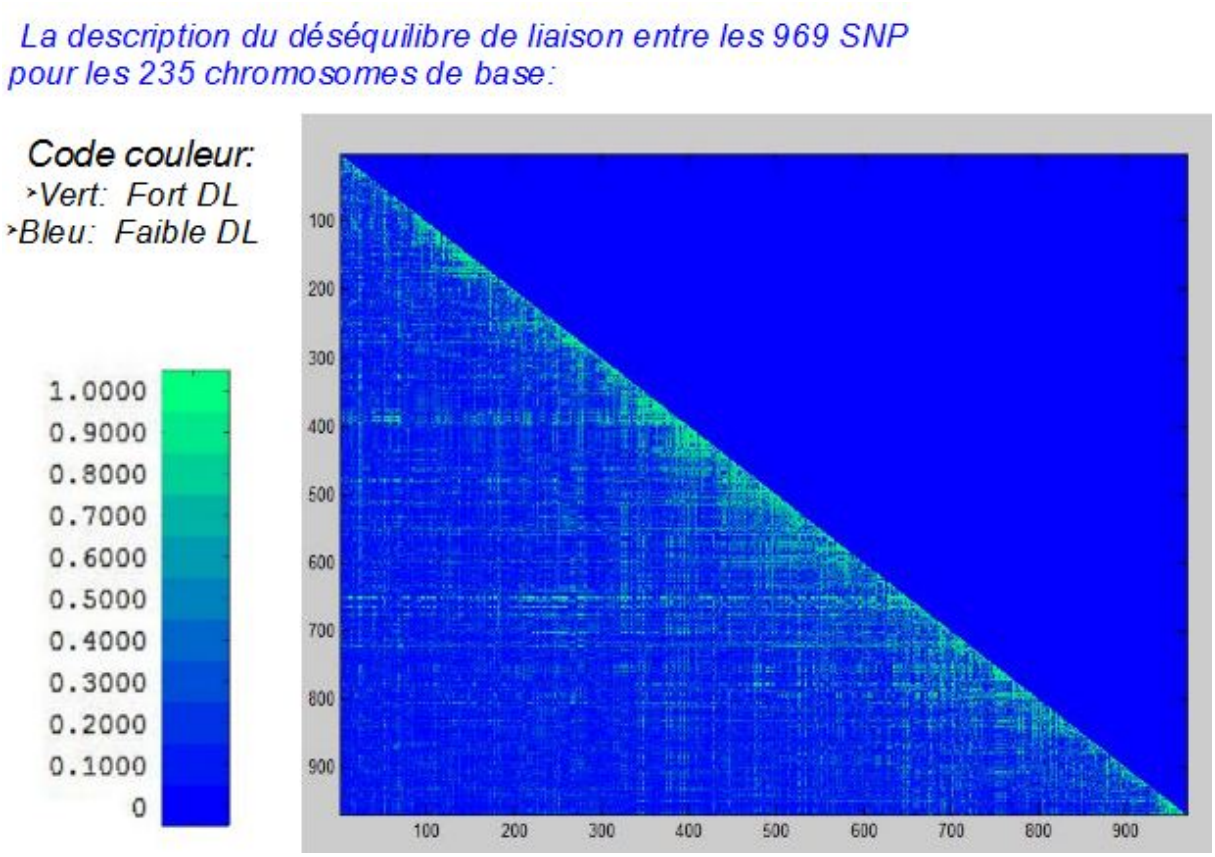

FIGURE 6 – **La description du DL entre les 969 SNP pour les 235 chromosomes de base**

Une autre représentation, où l'on fait glisser une fenêtre de 20 marqueurs, est donnée par la figure 7 ci-dessous. Pour chaque SNP de 1 à 950 on peut voir l'étendue du DL sur 20 marqueurs et la différence de cette étendue selon la mesure du D' et celle du r^2.

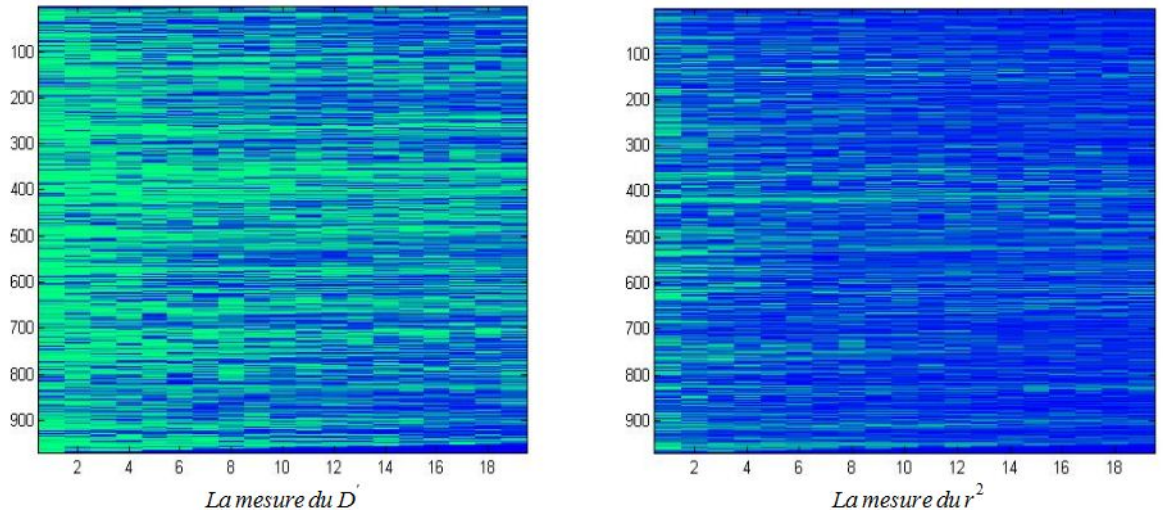

FIGURE 7 – **Etendue du DL sur 20 marqueurs selon les mesures pour les 235 chromosomes de base**

3.2 L'objectif des mesures de similitude (s(IBD)) entre segments chromosomiques

Les méthodes d'estimation de similitude entre chromosomes servent à produire des matrices de ressemblances entre des segments chromosomiques. L'information utilisée est la ressemblance entre les génotypes sur un ou plusieurs marqueurs successifs, 6 marqueurs dans notre étude. On compare ces matrices avec une matrice contenant les statuts IBD (0 ou 1) connus à un locus cible (QTL putatif). Le locus cible est un marqueur qui sera utilisé pour simuler un QTL comme évoqué précédemment. Comme on connaît l'origine chromosomique des allèles à tous les marqueurs on est capable de dire si l'allèle au marqueur cible pour un chromosome est IBD avec l'allèle au marqueur cible pour un autre chromosome. La figure 8 ci-dessous décrit cette situation, où Q représente l'allèle au QTL ayant un effet favorable sur le phénotype et q le contraire. Dans un modèle additif par exemple la combinaison d'allèles déterminant le génotype g_i au QTL déterminera l'effet sur le phénotype y_i d'un individu i, i.e $g_i = +2$ ou 0 ou-2 si l'individu possède les génotypes QQ ou Qq ou qq respectivement.

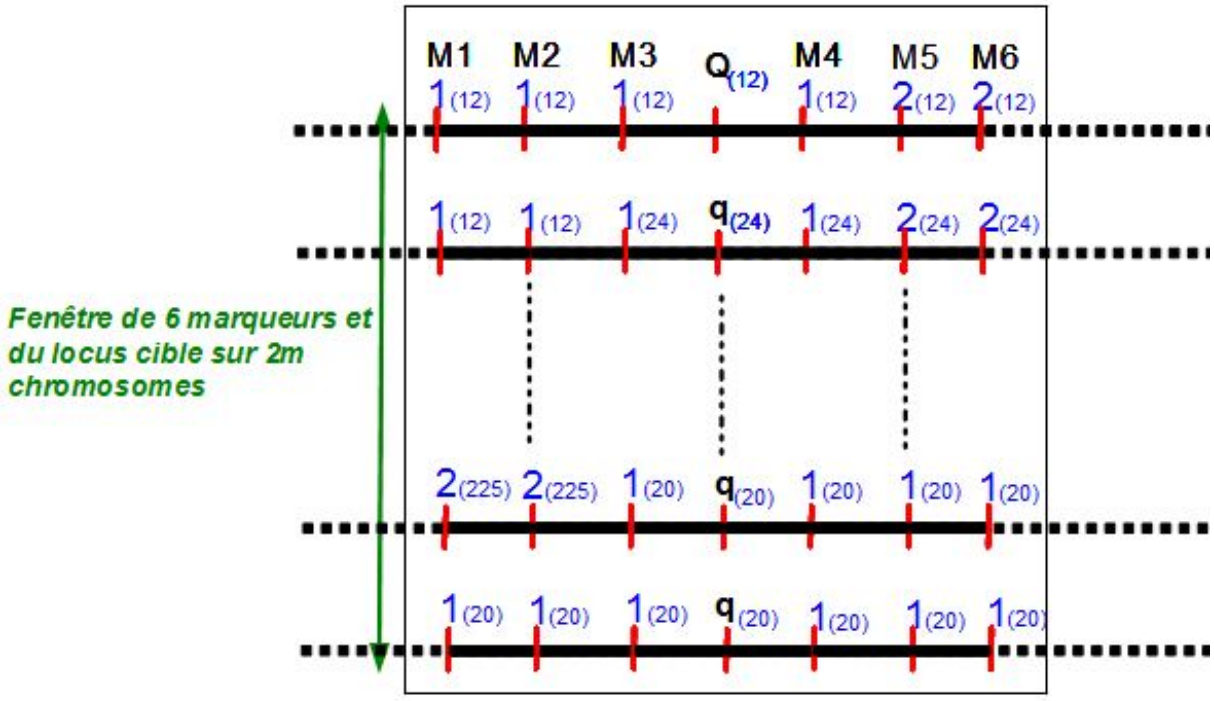

FIGURE 8 – **Fenêtre de 6 marqueurs et du locus cible sur les 2m=970 chromosomes des fondateurs pour la cartographie de QTL**

Sur une fenêtre donnée on a donc d'une part la comparaison sur la base de l'IBD au locus cible pour tous les chromosomes, deux à deux, qui permet de produire une matrice contenant des statuts IBD, et d'autre part les s(IBD) qui permettent de produire des matrices de similarité entre segments chromosomiques en ségrégation sur la fenêtre. Il est à noter que la notion d'haplotype fait référence à la séquence d'allèles pour les 6 marqueurs, alors que la notion de segments chromosomiques fait référence à l'échantillonnage parmi les haplotypes possibles.

3.2.1 Exemple d'un modèle mixte utilisant les matrices de similarité en cartographie de QTL

En utilisant l'information du DL que les s(IBD) captent on souhaite avoir un regroupement correct des segments chromosomiques ayant un allèle IBD au QTL, qu'on peut ensuite utiliser dans un modèle mixte pour tester la présence d'un QTL pour les haplotypes recensés à une fenêtre. Du fait de ce regroupement correct les segments chromosomiques porteront donc le même allèle, Q ou q, au QTL. Un exemple de modèle mixte pour la cartographie de QTL est le suivant :

$$Y = \mu + Wv + \varepsilon\,(H1) \qquad v/s \qquad Y = \mu + \varepsilon\,(H0),$$

tel que $v \sim \mathcal{N}(0, Z\sigma_v^2)$, $\varepsilon \sim \mathcal{N}(0, I\sigma_\varepsilon^2)$ et $-2ln(\frac{L0}{L1}) \sim \frac{1}{2}\chi(0)^2 + \frac{1}{2}\chi(1)^2$
où ;
- Y : Le vecteur des phénotypes (m x 1)
- μ : La moyenne générale partagée par les individus (m x 1)
- W : La matrice d'incidence reliant les effets haplotypiques aux individus (m x k)
- v : Le vecteur des effets aléatoires des haplotypes (k x 1)
- Z **: Matrice de ressemblance entre k x k haplotypes**
- ε : Le vecteur des résidus (m x 1)
- $L0$ et $L1$ sont les vraisemblances des données calculées sous $(H1)$ et $(H0)$ respectivement.
- $-2ln(\frac{L0}{L1})$ est le test du rapport de vraisemblance qui a pour loi $\frac{1}{2}\chi(0)^2 + \frac{1}{2}\chi(1)^2$

3.2.2 Les s(IBD) étudiées : l'IBS, le Score de Similarité et $\mathbb{P}$(IBD)

- L'IBS :
C'est la plus simple des mesures de similarité entre régions. Deux segments chromosomiques sont dits IBS s'ils partagent les mêmes allèles à tous les marqueurs sinon ils sont nonIBS. C'est une fonction booléenne, IBS$\rightarrow \{0, 1\}$.
Par exemple pour $h_1 = (122121)$, $h_2 = (122122)$ et $h_3 = (122121)$, on a IBS(h_1,h_2)=0 et IBS(h_1,h_3)=1.

- Le Score de Similarité (Li et Jiang, 2005) :
C'est une fonction somme qui se décompose en deux fonctions sommes dont la première compte le nombre d'allèles en commun entre deux segments chromosomiques et la deuxième calcule la longueur des segments partagés en commun par les deux segments chromosomiques. Le score entre deux segments chromosomiques h_i et h_j est donné par :

$$s_{i,j} = \sum_{k=-l}^{r} w_1(x_k)\mathbb{1}(h_i(k), h_j(k)) + \sum_{\substack{k=-l' \\ k\neq 0}}^{r'} w_2(x_k)$$

où x_k est la distance physique d'un marqueur $h(k)$, à la position k ($-l \leq k \leq r$, $-l' \geq -l$ et $r' \leq r$), par rapport au centre de l'haplotype, le centre est à la distance $x_0 = 0$. w_1 et w_2 sont des fonctions poids décroissantes qui ont pour objet de décroître l'importance que l'on attribue aux marqueurs quand ceux-ci s'éloignent du locus de référence x_0. Dans le cadre de mon stage je les ai définies comme $w_1(x_k) = w_2(x_k) = 1 - \sum_{i=0}^{k-1}(x_i - x_{i+1})$ où $(x_i - x_{i+1})$ est la longueur réelle en centiMorgan entre les marqueurs situés aux distances x_i

et x_{i+1} par rapport au locus de référence x_0. Par exemple si $\forall i \in \{-l, .., r\}$ $(x_i - x_{i+1}) = 0,1$ (distance constante entre les marqueurs) alors pour deux segments chromosomiques de 6 marqueurs, $h_1 = (12\mathbf{21}22)$ et $h_2 = (12\mathbf{21}21)$ centrés à mi-chemin entre l'allèle **2** et **1**, on a :

$$\sum_{k=-3}^{3} w_1(x_k)\mathbb{1}(h_1(k), h_2(k)) = 2\left[1 - \frac{0,1}{2}\right] + 2\left[1 - (0,1 + \frac{0,1}{2})\right] + \left[1 - (0,1 + 0,1 + \frac{0,1}{2})\right] = 4,35$$

$$\text{et } \sum_{\substack{k=-3 \\ k\neq 0}}^{3} w_2(x_k) = [1 - 0,1] + 2\left[1 - (0,1 + \frac{0,1}{2})\right] + \left[1 - (0,1 + 0,1 + \frac{0,1}{2})\right] = 3,35$$

$$\Longrightarrow s_{1,2} = 4,35 + 3,35 = 7,7$$

De la même façon on obtient le score maximum entre deux segments chromosomiques de 6 marqueurs, i.e : $s_{max} = s_{1,1} = s_{2,2} = s_{i,i} = 5,1 + 4,1 = 9,2$.
On a donc que le score normalisé entre h_1 et h_2 vaut : $s_{1,2}^{norm} = \dfrac{s_{1,2}}{s_{max}} = 0,8370$

La construction de ce score comme mesure de similarité vient du fait que deux segments chromosomiques partageant un plus long segment en commun, ou une somme plus importante de segments en commun à droite et à gauche du locus de référence, ont plus de chance d'avoir été hérités d'un ancêtre en commun après des évènements de recombinaisons, ce qui est capté par la deuxième somme. La première somme quant à elle sert à évaluer le degré de ressemblance qu'il peut y avoir entre deux segments chromosomiques même s'ils ne partagent éventuellement pas de segment en commun. Cette première fonction somme évite aussi qu'on sous-estime le degré de ressemblance qu'il pourrait y avoir si jamais il se produit une mutation ou une erreur de génotypage par rapport à l'un des marqueurs d'un haplotype. L'exemple suivant illustre le concept de ce score de similarité (sans considération des fonctions poids cette fois-ci) :
Si l'on considère les 4 segments chromosomiques $h1 = (\mathbf{112}1\mathbf{2}1)$, $h2 = (\mathbf{122}2\mathbf{2}2)$, $h3 = (1\mathbf{122}12)$ et $h4 = (2\mathbf{122}21)$ et que l'on considère seulement le plus long segment entre deux segments chromosomiques comme mesure de similarité alors $s(h_3, h_4) = 2$ tandis que $s(h_1, h_2) = 0$ alors que h_1 et h_2 partagent 3 allèles en commun. De plus si on a une erreur de génotypage, ou une mutation, au 2-ième marqueur de h_1 ou de h_2 alors la deuxième fonction somme seulement sous-estimerait vraiment la ressemblance entre ces segments chromosomiques. Les deux mesures combinées définissent un score bien plus robuste que chacune des mesures prise séparément.
Pour ce score de similarité (normalisé) on peut définir la mesure du score seuillé où l'on doit dépasser un certain seuil, qui est une mesure de probabilité en soi, afin d'avoir une probabilité importante d'être IBD ou la probabilité d'être IBD est de 0 sinon. Par exemple pour le score de similarité seuillé à α on a que si $s_{i,j} < \alpha$ alors $s_{i,j} = 0$, avec $\alpha \in [0,6; 1[$.

- $\mathbb{P}$(IBD) (Meuwissen & Goddard, 2001) :
$\mathbb{P}$(IBD) cacule la probabilité qu'un allèle à un locus A est IBD conditionnellement à l'état IBS ($S = 1$ ou 0) des marqueurs adjacents. Cette mesure intègre les hypothèses d'un modèle de coalescence sous-jacent, analogue à celui de Wright-Fisher, dans ce calcul de probabilité. Par exemple la probabilité que deux segments chromosomiques $h_1 = (122 \underset{locus A}{\mathrm{H}} 122)$ et $h_2 = (122 \underset{locus A}{\mathrm{H}} 121)$ aient des allèles IBD au centre (le locus A) est donnée

par :

$$\mathbb{P}(A = IBD|S) = \frac{\mathbb{P}(A = IBD\&S)}{\mathbb{P}(S\&A = IBD) + \mathbb{P}(S\&A = nonIBD)}$$

$\underset{locusA}{\mathrm{H}}$ est l'allèle au locus A qu'on n'observe pas et dont on veut calculer la probabilité d'être IBD pour les deux segments chromosomiques, $\underset{locusA}{\mathrm{H}}$ correspond donc au locus cible dans les simulations et au QTL putatif dans la cartographie

Afin d'effectuer ce calcul les auteurs définissent une variable ϕ, qui s'apparente à un vecteur, qui contient les différents états d'IBD (0 ou 1) des allèles des deux segments chromosomiques et aussi l'information de recombinaison historique, éventuel, entre chacun de ces allèles. Par exemple pour des haplotypes de 3 marqueurs, donc de 3 allèles, le vecteur : $\phi = [\phi(-2)\ \ \phi(-1)\ \ \phi(0)\ \ \phi(1)\ \ \phi(2)] = [1\_1 \times 0]$, décrit la situation où les allèles aux positions -2 et 0 sont IBD ($\phi(-2) = 1$ & $\phi(0) = 1$) sans aucun évènement de recombinaison au milieu ($\phi(-1) =$ "_") et l'allèle à droite du locus 0 est nonIBD ($\phi(2) = 0$) précédé d'une recombinaison ($\phi(1) =$ " $\times$ "). On remarquera qu'un statut nonIBD est forcément toujours précédé ou suivi d'une recombinaison. Cette description se généralise à tout haplotype de plusieurs marqueurs.
Si on suppose que le locus A est à la position 0, ce qui est toujours le cas, alors on a : $\mathbb{P}(S\&A = IBD) = \sum_{\phi|\phi(0)=1} \mathbb{P}(S|\phi)\mathbb{P}(\phi)$ et $\mathbb{P}(S\&A = nonIBD) = \sum_{\phi|\phi(0)=0} \mathbb{P}(S|\phi)\mathbb{P}(\phi)$

$\mathbb{P}(\phi)$ est calculée selon les hypothèses d'un modèle de coalescence à savoir que la probabilité qu'il n'y ait pas d'ancêtre commun à deux gamètes pendant $t-1$ générations est de $\left(1 - \frac{1}{2Ne}\right)^{t-1}$, où $2Ne$ est la taille de la population des gamètes pour des organismes diploïdes et $t \in \{1, .., T\}$ tel que $T =$ génération actuelle. De plus les auteurs imposent que la probabilité qu'il n'y ait pas de recombinaison entre deux loci vaut $exp(-c)$ (distribution de Poisson de k=1 occurrence dans un intervalle de longueur c) où c est la longueur en Morgan. Par exemple la probabilité qu'une région de deux marqueurs successifs (incluant les marqueurs) soit complètement IBD, donc qu'on n'ait aucune recombinaison depuis la 1ère génération $t = 1$ jusqu'à la génération actuelle $t = T$, est donnée par :

$$\mathbb{P}(\phi = [1\_1]) = \frac{1}{2Ne} exp(-2c) \sum_{t=1}^{T} exp\left[-(t-1)\left(\frac{1}{2Ne} + 2c\right)\right]$$

De ce calcul les auteurs dérivent des expressions pour le calcul de la probabilité de ϕ selon les recombinaisons. Ces calculs se généralisent à tout haplotype de plusieurs marqueurs. On pourra se reporter à l'article pour plus de détails et d'exemples.

Durant mon stage j'ai eu à programmer en Fortran 90, dans la plus grande généricité, les deux premières s(IBD) qui sont l'IBS et le score de similarité. Toutefois j'ai pu bénéficier du logiciel existant pour le calcul de $\mathbb{P}$(IBD), dont le code en Fortran 90 est fourni dans l'article de Meuwissen & Goddard, 2001.

4 La comparaison des s(IBD) avec les statuts IBD connus

4.1 La comparaison sur la base des $2m \times 2m$ chromosomes fondateurs

A une fenêtre il peut arriver qu'on ait deux segments chromosomiques qui soient IBS tandis que le marqueur cible simulant le QTL n'est pas IBD pour ces deux segments chromosomiques. On rappelle que deux allèles qui sont IBD sont forcément IBS (on ne considère pas de mutation dans les simulations) tandis que le fait d'être IBS n'implique pas le fait d'être IBD. En ce sens on aura donc deux classes pour l'origine IBD au marqueur cible et une seule classe pour l'IBS, pour ces deux segments. La figure 9 suivante décrit cette situation, en considérant seulement 2 segments chromosomiques à une fenêtre, où l'on a deux classes pour l'origine IBD au marqueur cible (l'allèle Q au QTL provenant du chromosome 12 et q du chromosome 24) et une seule classe d'haplotype(**111122**) sur la base de l'IBS.

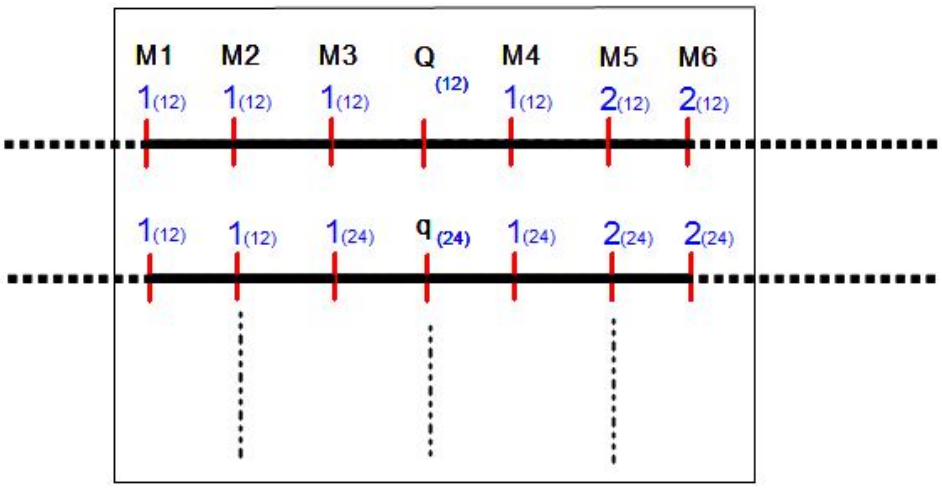

FIGURE 9 – **Deux classes pour l'IBD et une classe pour l'IBS pour deux segments chromosomiques**

Cela exprime le fait qu'il est quasiment impossible de regrouper les segments chromosomiques selon leur statut IBS, et les allèles au QTL selon leur origine IBD, en un même nombre de classes. Ce problème est encore accentué si l'on considère que l'on a $2m = 970$ segments chromosomiques à une fenêtre. Comme les marqueurs sont bialléliques on peut donc avoir k classes, $k \in \{1, 2, .., 2^6\}$, d'haplotypes possibles et l classes, $l \in \{1, 2, .., 235\}$, pour l'origine IBD de l'allèle au QTL à une fenêtre de $2m$ segments chromosomiques en ségrégation. A une fenêtre il est donc impossible de comparer directement les matrices $k \times k$ de similarités avec une matrice $l \times l$ contenant les représentants de chacune des classes de l'origine IBD de l'allèle au QTL.
De ce fait, pour la comparaison des méthodes de similarité avec les statuts IBD connus, on utilise les matrices de mesure de similitude entre les $2m \times 2m$ segments chromosomiques

et la matrice $2m \times 2m$ contenant les status IBD au QTL. Dans la suite on référa à ces matrices $2m \times 2m$ en tant que IBS, $SCORE$, $SCORE\_S$ (la matrice de la méthode du score seuillé à 0,8) et $PrIBD$, en référence aux mesures qui les produisent, et à IBD pour la matrice contenant les status IBD au QTL. On a donc la définition suivante : soit E_s l'ensemble des matrices symétriques réelles produites par les s(IBD) étudiées à une fenêtre, $\boldsymbol{E_s = \{IBS, SCORE, SCORE\_S, PrIBD\}}$.

4.2 La définition des vrais et faux positifs

Soit $M \in E_s$ tel que $M = IBS, SCORE\_S$ ou $PrIBD$. On pose $M = (m_{ij})_{1 \leq i,j \leq 2m}$ et $IBD = (b_{ij})_{1 \leq i,j \leq 2m}$. On appelle les vrais positifs de la matrice M les éléments $m_{i'j'} \geq \beta$ tels que $b_{i'j'} = 1$, où $\beta = inf\{m_{i,j} \neq 0;\ 1 \leq i, j \leq 2m\}$. Pour $M = IBS$ on a $\beta = 1$, pour $M = PrIBD$ on a $\beta \geq 0,9996$ et pour $M = SCORE\_S$ on a $\beta = 0,8$. Les faux positifs de la matrice M sont quant à eux les éléments $m_{i'j'} \geq \beta$ tels que $b_{i'j'} = 0$.

Il est à noter que les éléments non nuls de IBD détectés par une mesure de similarité continue tel que le score de similarité ne pourront pas être considérés comme des vrais, et faux positifs, si on n'impose aucun seuil d'attribution d'avoir une probabilité importante d'être IBD pour ces derniers, d'où le fait de seuiller la matrice $SCORE$ à 0,8 afin d'obtenir la matrice $SCORE\_S$ par exemple.

4.3 La comparaison des matrices de similarité avec IBD

4.3.1 La nature des matrices de similarité et de IBD

$\forall M \in E_s$ on observe que :

.***M* est presque toujours non semi-définie positive** (les valeurs propres ne sont pas toutes positives) .
.M est sparse et a des attributs binaires pour $M = IBS$ (0 ou 1), $PrIBD$ (0 ou 1, 0 ou 0,9997 par exemple)

A une fenêtre on a aussi que :

.***IBD* est presque toujours non semi-définie positive**.
.IBD est sparse et a des attributs binaires (0 ou 1)

Comparer M à IBD n'est pas une tâche facile car il n'existe pas de théorie générale sur la comparaison des matrices, et l'ensemble des matrices carrées n'est muni que d'une relation d'ordre partielle compatible avec l'addition et la multiplication. Une étude sur la comparaison de matrices décrite dans l'article de Schneider & Borlund (2007,2008) spécifie qu'il existe quelques théories sur le sujet mais que cela dépend surtout de la nature des matrices et qu'en conséquence les mesures appropriées pour la comparaison sont très souvent définies en fonction de ces dernières.
Une méthode de comparaison graphique de matrices de variance-covariance considérées comme des matrices de similarités $((M_i)_{i \in \{1,..,k\}})$ a été développée par Escoufier et L'Hermier (1978). Cette étude est basée sur des concepts "d'interstructure" et "d'intrastructure" qui s'appuient sur le calcul des valeurs propres d'une matrice S qui est tel que $(s_{i,j})_{i,j \in \{1,..k\}} = Tr(M_i M_j) = COVV(M_i, M_j)$ (notation d'Escoufier) et le calcul des

valeurs propres des $(M_i)_{i\in\{1,..,k\}}$. Toutefois une condition nécessaire pour utiliser cette méthode est que les $(M_i)_{i\in\{1,..,k\}}$ soient semi-définies positives ce qui n'est quasiment jamais vrai pour M dans notre cas.

4.3.2 Les mesures de comparaison

Comme les matrices de similarité produites (E_s) sont sparses, et non semi-définie positive, il est nécessaire de définir des mesures adaptées afin de pouvoir les comparer avec la matrice IBD. En ce sens on peut vouloir calculer une mesure de distance entre la matrice M et IBD afin de trouver la mesure de similitude qui se rapproche au mieux d'IBD. Toutefois, même si nos objets matrices sont proches en terme de distance, il n'en demeure pas moins que la mesure (ou une combinaison de mesures) utilisée devra être suffisamment fiable pour décrire la ressemblance entre ces derniers. Il est à noter que toutes les matrices produites par les s(IBD) ont été normalisées avant toute comparaison avec IBD.

Soit E l'espace vectoriel sur $\mathbb{R}$ définit par l'ensemble des matrices réelles symétriques. L'ensemble E est canoniquement muni d'une structure d'espace euclidien, ce qui peut nous conduire à calculer la distance euclidienne entre $M = (m_{ij})_{1\leq i,j\leq 2m}$ et $IBD = (b_{ij})_{1\leq i,j\leq 2m}$ définie de la façon suivante :

$$\textbf{1. } d(M, IBD) = \| M - IBD \| = \left[\sum_{i=1}^{2m}\sum_{j=1}^{2m}(m_{ij} - b_{ij})^2\right]^{\frac{1}{2}}$$

On peut aussi définir sur E ;

Un produit scalaire entre matrices qui correspond à la trace du produit des matrices :

$$\textbf{2. } \forall(M_1, M_2) \in E \times E : \langle M_1, M_2\rangle = tr(M_1 M_2)$$

La norme associée (dite de Frobenius) :

$$\textbf{3. } \forall M \in E : \| M \| = \sqrt{tr(M^T M)} = \left(\sum_{1\leq i,j\leq 2m} m_{i,j}^2\right)^{\frac{1}{2}}$$

L'angle entre deux éléments de E :

$$\textbf{4. } \theta_{M_1,M_2} = \arccos\left(\frac{tr(M_1 M_2)}{\sqrt{(tr(M_1^2)tr(M_2^2))}}\right) \quad ; \theta \in [0, \pi]$$

Toutefois le simple fait d'être proche (ou éloigné) avec la distance euclidienne n'assure pas qu'on ait des objets qui se ressemblent (ou non) réellement car les matrices de similarité sont de normes différentes. La distance n'est donc pas une mesure appropriée car elle ne contient pas toute l'information de la dissimilarité existante entre les matrices. Par exemple soit $M_1, M_2 \in E_s$ de normes différentes. La figure 10 ci-dessous décrit une situation pour M_1 et M_2, où M_1 est à une plus grande distance d'IBD que M_2, mais en n'étant pas forcément d'avantage dissimilaire à IBD que M_2.

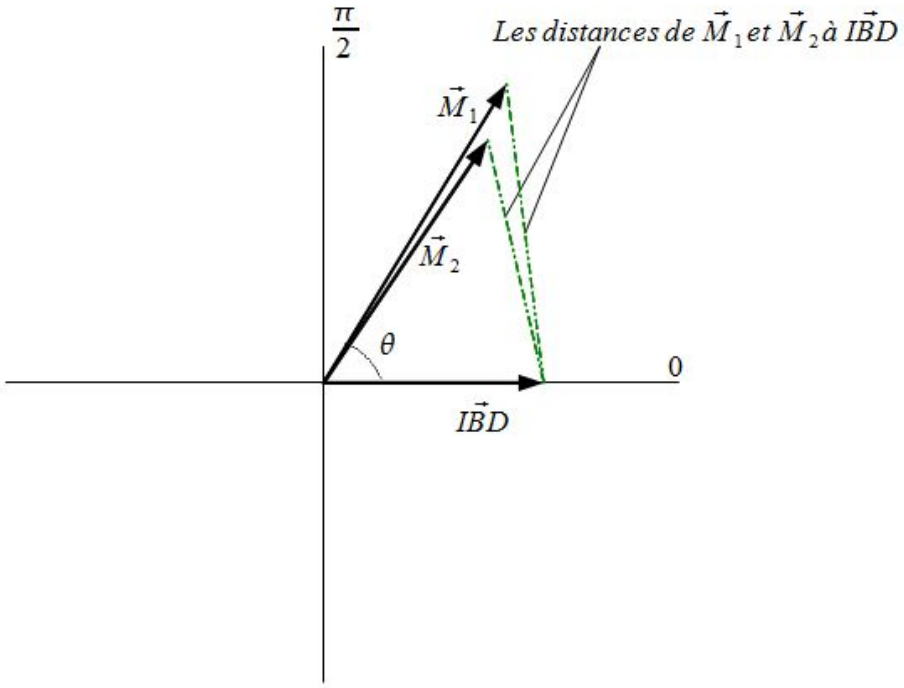

FIGURE 10 – **Les distances de M_1 et M_2 à IBD**

Cependant on remarque que la projection de M_1 sur IBD ($P_{IBD}M_1$) a un impact direct sur le produit scalaire entre ces derniers, ce qui est intrinsèquement lié à la définition même du produit scalaire usuel entre deux vecteurs dans un espace euclidien ;
i.e : $\langle M_1, IBD\rangle = \| P_{IBD}M_1 \| \| IBD \|$ (idem pour M_2 sur IBD). Le produit scalaire entre deux vecteurs dans un espace euclidien donne une mesure de covariance empirique entre ces deux vecteurs et en normalisant ce produit scalaire par la norme d'IBD au carré ; i.e : $\| IBD \|^2 = \langle IBD, IBD\rangle$, on ramène les projections au cercle unité. La figure 11 ci-dessous montre la projection de M_1 et de M_2 sur IBD :

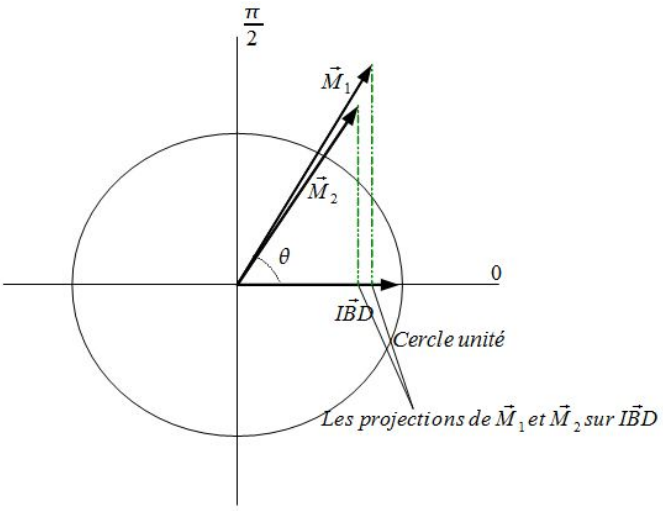

FIGURE 11 – **Les projections de M_1 et M_2 sur IBD**

4.3.3 Le choix et l'interprétation de la mesure de covariance entre les matrices : la statistique de Mantel

Pour le calcul du produit scalaire normalisé on n'utilisera pas la trace du produit matricielle définie précédemment mais le produit scalaire défini entre la partie triangulaire supérieure de M et celle de IBD. Ce produit scalaire est connu sous le nom de la statistique de Mantel ;

$$\text{i.e :} \quad z_{M,IBD} = \sum_{1 \leq i < j \leq 2m} m_{ij} b_{ij}.$$

Ce choix a été fait pour deux raisons. La première est que les valeurs empiriques associées sont pratiquement les mêmes et que le temps de calcul associé est nettement inférieur à celui de la trace du produit matriciel. Le fait que les valeurs empiriques sont pratiquement les mêmes se "démontre" dans une certaine mesure. En effet si on considère la norme de Frobenius de la matrice IBD alors on sait que :

$$\| IBD \|^2 = tr(IBD^T IBD) = \sum_{1 \leq i,j \leq 2m} b_{i,j}^2$$

Or, soit $N = (m_{ij})_{1 \leq i < j \leq 2m}$ le vecteur correspondant à la partie triangulaire supérieure de M et $B = (b_{ij})_{1 \leq i < j \leq 2m}$ celui de IBD, on a :

$$\langle B, B \rangle = \sum_{1 \leq i < j \leq 2m} b_{ij}^2 = z_{IBD,IBD} = K_1 tr(IBD^T IBD)$$

$$\text{et } \langle N, B \rangle = \sum_{1 \leq i < j \leq 2m} m_{ij} b_{ij} = z_{M,IBD} = K_2 tr(M^T IBD) \text{ tel que } K_2 = K_1 + \epsilon$$

,où $K_1, K_2 \in \mathbb{R}$ et $\epsilon = o(K_1)$.

Donc on a bien que : $z_{M,IBD}^{norm} = \dfrac{z_{M,IBD}}{z_{IBD,IBD}} = \dfrac{\langle N, B \rangle}{\langle B, B \rangle} = \dfrac{K_2 tr(M^T IBD)}{K_1 tr(IBD^T IBD)} \approx \dfrac{tr(M^T IBD)}{tr(IBD^T IBD)}$

La deuxième raison est que toute l'information de similarité se trouvent sur la partie triangulaire supérieure excluant la diagonale (ou inférieure excluant la diagonale car ce sont les mêmes) de M. Donc cela revient aussi à travailler dans un espace euclidien de dimension $2m(2m-1)/2$ (où $2m(2m-1)/2$ est le nombre d'éléments de la partie triangulaire supérieure excluant la diagonale) et à appréhender le concept de ressemblance entre objets autrement. C'est à dire entre vecteurs dans un espace euclidien $\mathbb{R}^{2m(2m-1)/2}$, au lieu de matrices, car ce sont les mêmes objets que l'on compare dans le cadre d'espaces vectoriels.

Comme les matrices M et IBD et les vecteurs N et B associés respectivement à ces matrices sont sparses, il vient que $z_{M,IBD}$ donne une mesure de comptage , ou de détection ("covariance"), des éléments non nuls de B par N. Autrement dit $z_{M,IBD}$ donne une mesure de comptage des éléments non nuls (les 1) de IBD détectés par des éléments non nul (pas toujours égaux à 1) de la matrice M produite par une s(IBD). Toutefois on ne peut pas parler d'une réelle covariance entre M et IBD car les zéros de la matrice M et de IBD sont des éléments absorbants dans le produit scalaire de N avec B. Par exemple si M est produite par l'IBS , donc N aussi, on a ;

$$z_{M,IBD} = \langle N, B \rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ . \\ . \\ 1_{i-ième} \\ 0 \\ . \\ . \\ 1 \\ 1 \\ 1_{n-ième} \end{pmatrix}_N \begin{pmatrix} 0 \\ 1 \\ 0 \\ . \\ . \\ 0_{i-ième} \\ 1 \\ . \\ . \\ 0 \\ 0 \\ 1_{n-ième} \end{pmatrix}_B$$

et on remarque qu'il y a des statuts IBD à 1 du vecteur B qui sont détectés par les statuts IBS à 1 du vecteur N (des vrais positifs) mais qu'on a aussi des statuts IBD à 0 du vecteur B qui sont détectés à 1 par le vecteur N (des faux positifs) qu'on ne somme pas dans le produit scalaire entre B et N, il en va de même pour les vrais et faux négatifs. Donc la statistique de Mantel seulement ne suffit pas à décrire la ressemblance entre les objets matrices et il nous faut donc une autre mesure combinée à cette statistique afin de décrire au mieux cette ressemblance.

Pour obtenir le niveau de signification (la p-value) associé à l'hypothèse nulle ($H0$) d'une non corrélation entre les éléments de M et de IBD on utilise un test de permutation où l'on permute les éléments de N en gardant ceux de B fixe. On calcule d'abord la statistique originale $z_{M,IBD}$ sans permutation de N. Ensuite pour toute nouvelle permutation de N la statistique $z_{M,IBD}$ est calculée à nouveau, cela produit la distribution empirique sous l'hypothèse nulle de référence qui exprime le fait qu'il n'existe pas de corrélation entre les éléments de M et de IBD. La p-value ensuite calculée représente la probabilité qu'une des statistiques calculées pour la distribution nulle de référence est plus grande ou égale à la statistique originale. Si elle est petite alors on rejette $H0$ et on conclue à une corrélation entre les 2 vecteurs donc entre les 2 matrices (il existe un lien fort entre les éléments des 2 matrices), sinon la statistique originale semble aléatoire et appartenir à la distribution nulle de référence.

4.3.4 L'homothétie comme mesure de ressemblance dans un espace euclidien

Comme il a été dit précédemment la statistique $z_{M,IBD}$ sert surtout à donner une mesure des éléments non nuls de IBD détectés par ceux de M. Elle décrit donc difficilement la ressemblance qu'il pourrait y avoir entre M et IBD. Comme on cherche à se rapprocher de IBD par une matrice qui possède les mêmes propriétés que IBD il semble plus approprié de définir une autre mesure qui puisse évaluer la ressemblance entre M et IBD en terme de propriétés. Soit $\alpha \in \mathbb{R}^*$. On appelle homothétie vectorielle de rapport α l'application h_α définie de $\mathbb{R}^n$ dans $\mathbb{R}^n$ qui est telle que $\forall N \in \mathbb{R}^n \; h_\alpha(N) = \alpha N$. L'idée d'homothétie comme mesure de ressemblance vient du fait que deux objets sont semblables si l'un est le dilaté de l'autre par un scalaire et qu'être semblable dans un espace euclidien $\mathbb{R}^n$ c'est aussi avoir les mêmes propriétés structurelles. Par exemple, les rapports des mêmes côtés pour deux objets en homothétie sont toujours égaux.

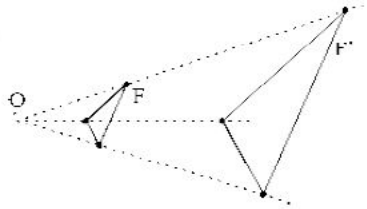

FIGURE 12 – **Une homothétie dans $\mathbb{R}^n$**

On choisit donc la mesure d'angle, ou la distance géodésique, sur le cercle unité, entre les vecteurs N et B, comme étant l'une des mesures qui servira à décrire la ressemblance entre M et IBD. Plus M sera proche d'IBD en termes d'angle et plus M sera proche d'IBD en terme de structure matricielle et donc de propriétés. Les seuls éléments de la matrice M qui ont pour effet de la rendre dissemblable à la matrice IBD sont ceux ayant un état de vérité faux dans la capture des statuts (non)IBD. La mesure d'angle est donc une fonction croissante de ces éléments. Toutefois la condition de moindre angle entre M et IBD est une condition non suffisante d'optimalité (sauf pour $\theta_{M,IBD} = 0$) pour la ressemblance, même si elle est nécessaire. Il peut arriver par exemple qu'on ait deux matrices $M_1, M_2 \in E_s$ qui ont un même angle par rapport à l'IBD, mais dont la première M_1 a un plus grand produit scalaire avec l'IBD. La figure 12 ci-dessous illustre cette situation.

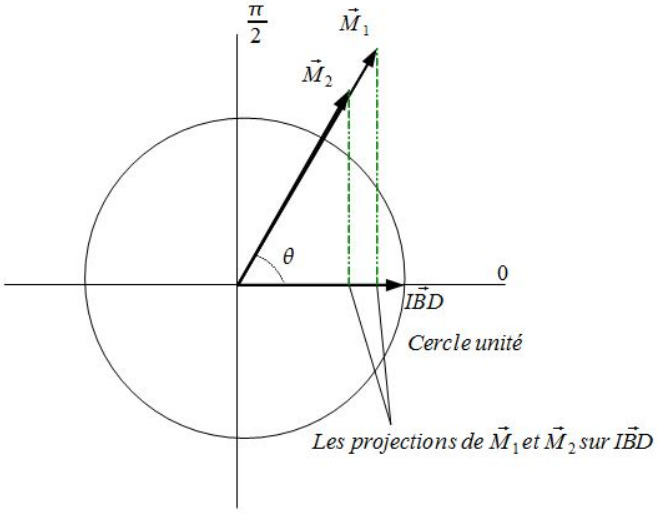

FIGURE 13 – **Les projections de $\vec{M}_1$ et de $\vec{M}_2$ (qui sont en homothétie) sur $I\vec{B}D$**

Dans ce cas particulier il vient donc que la matrice M_1 est un meilleur estimateur de la matrice IBD que la matrice M_2. Un estimateur naturel d'IBD à une fenêtre pourrait alors être défini comme ci-suit :

$$\hat{IBD} = \underset{M \in E_s}{argmax}\ z_{M,IBD} \cap \underset{M \in E_s}{argmin}\ \theta_{M,IBD}$$

4.3.5 Le théorème de projection dans un espace de Hilbert

Définition : Un espace de Hilbert H est un espace vectoriel normé complet dont la norme associé $\| . \|$ provient d'un produit scalaire ou hermitien $< ., . >$.

Remarque : Un espace de Hilbert est la généralisation en dimension quelconque d'un espace euclidien ou hermitien et un espace est dit complet si toute suite de Cauchy a sa limite dans l'espace.

Théorème :
Soit H un espace de Hilbert, et K un sous espace vectoriel fermé non vide de H. Soit $x \in H$ alors $\exists!\ y^* \in K$ qui vérifie : $\| x - y^* \|_H = \inf_{y \in K} \| x - y \|_H$ **(1)**

Caractérisation de y* : y^* est l'unique élément de K qui vérifie :

$$\forall y \in K \ \langle x, y \rangle = \langle y^*, y \rangle \qquad \textbf{(2)}$$

En effet on a $x = y^* + (x - y^*)$ et par linéarité du produit scalaire ;

$$\langle x, y \rangle = \langle y^*, y \rangle + \langle x - y^*, y \rangle \quad où \quad \langle x - y^*, y \rangle = 0$$

Remarque, on a aussi :

$$\forall y \in K, \langle x - y^*, y \rangle = 0 \qquad \textbf{(3)}$$

La figure 14 ci-dessous montre la projection de $x \in H$ sur $K \subset H$.

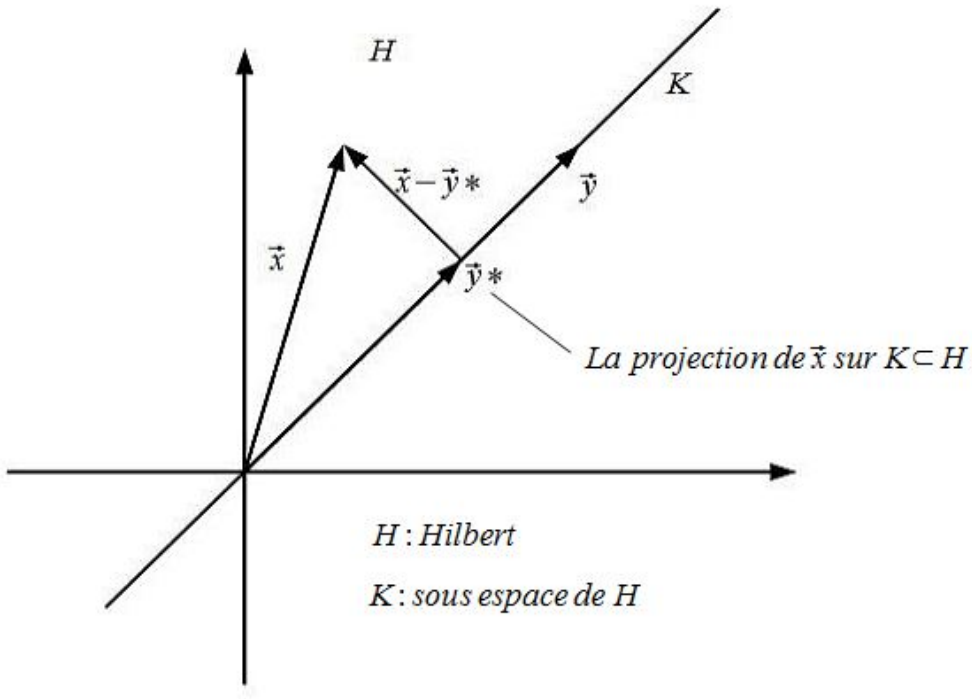

FIGURE 14 – **La projection dans un espace de Hilbert**

y^* est aussi la meilleure explication linéaire de x par les éléments de K et c'est celui qui minimise la "Mean Square Error" (i.e MSE= $\mathbb{E}[(\hat{x}-x)^2]$ est minimale pour $\hat{x}=y^*$) pour les éléments $y \in K$.

E est un Hilbert et E_s est un sous-espace vectoriel fermé non vide de E donc si on pose $x = IBD$ et $y = M$ alors la caractérisation (3) de y^* nous conduit à chercher l'élément $M^* \in E_s$ qui est tel que ;

$\forall M \in E_s,\ \langle IBD - M^*, M\rangle = 0$
$\Rightarrow\ \langle IBD - M^*, M^*\rangle = 0.$
Donc $\langle \boldsymbol{IBD - M, M}\rangle$ donne une mesure de comparaison pour la matrice $\boldsymbol{M \in E_s}$ avec $\boldsymbol{IBD}$, et le meilleur estimateur d' $\boldsymbol{IBD}$ au sens de la projection dans un Hilbert est atteint pour $\boldsymbol{M = M^*}$.

Cependant pour une matrice M qui possède que des attributs binaires, qui sont des 0 et des 1, on remarque que l'expression $\langle B - N, N\rangle$ donne une mesure de comptage négative des faux positifs de la matrice M par rapport à IBD. En effet pour le couple (i, j) fixé on a les 4 possibilités suivantes pour le produit de $(b_{ij} - m_{ij}).m_{ij}$: (i) $(1-1).1 = 0$, (ii) $(0-1).1 = -1$ (un faux positif compté négativement), (iii) $(1-0).0 = 0$ et (iv) $(0-0).0 = 0$.

Pour une matrice M qui a un ensemble d'attributs (qui possède un nombre d'éléments fini) inclus dans l'ensemble $\{0\} \cup [\alpha; 1]$ il existe un grand nombre de cas possibles pour le produit de $(b_{ij} - m_{ij}).m_{ij}$ à i et j fixés. Par simplicité on regarde seulement les 4 cas possibles suivants pour l'ensemble $\{0, \alpha\}$, avec $\alpha = 0,8$ par exemple, pour (i, j) fixé. Les 4 cas possibles sont : (i) $(1 - 0,8).0,8 = 0,16$,(ii) $(0 - 0,8).0,8 = -0,64$ (un faux positif compté négativement et pondéré par lui même), (iii) $(1-0).0 = 0$ et (iv) $(0-0).0 = 0$.

On remarque que l'on n'obtient pas les mêmes résultats pour le calcul de $\langle B - N, N\rangle$ selon si M a des attributs appartenant à l'ensemble $\{0, 1\}$ ou un ensemble d'attributs inclus dans l'ensemble $\{0\} \cup [\alpha; 1]$. Toutefois pour α proche ou égale à 1 ($\alpha \geq 0,8$ dans la pratique) on remarque qu'un faux positif de la matrice M a une contribution bien plus importante, relative aux 3 autres états de vérités, dans l'éloignement du résultat de $\langle B - N, N\rangle$ par rapport à 0.

4.3.6 Les mesures de comparaison et la nature des matrices

On donne un exemple concret en se plaçant en dimension $2m(2m-1)/2 = 3$ par simplicité. On pose $B = \mathbf{(1, 1, 0)}$, $N_1 = (1, \mathbf{0, 0})$, $N_2 = (1, 0, 8, 0)$ et $N_3 = (1, 0, 8, 0, 8)$. On remarque d'abord que $B \in Vect\{N_1, N_2\}$ et qu'il vient donc que le vecteur N^* associé à M^* (la caractérisation (3)) est donné par $N^* = B$. Les mesures de comparaison donnent les résultats suivants :

$$\begin{cases} \langle N_1, B\rangle = 1,\ \theta_{N_1,B} = \dfrac{\pi}{4} \approx 0,785 \text{ et } \mathrm{abs}(\langle B - N_1, N_1\rangle) = 0 \\ \langle N_2, B\rangle = 1,8,\ \theta_{N_2,B} \approx 0,111 \text{ et } \mathrm{abs}(\langle B - N_2, N_2\rangle) = 0,16 \\ \langle N_3, B\rangle = 1,8,\ \theta_{N_3,B} \approx 0,568 \text{ et } \mathrm{abs}(\langle B - N_3, N_3\rangle) = 0,48 \end{cases}$$

On voit que N_2 est d'avantage similaire avec B que N_1 avec ce dernier, ce qui n'est pas exprimé par la mesure de la caractérisation (3) pour N_1 à cause du faux négatif de ce dernier. Le fait que N_3 soit plus dissimilaire de B que N_2 est quant à lui exprimé car N_3

n'a aucun faux négatif. Le résultat de $\langle B - N, N\rangle$ est donc affecté par certains 0, les faux négatifs, de la matrice M associé au vecteur N. $I\hat{B}D$ semble dans ce cas plus fiable dans la comparaison de M avec IBD et la mesure d'angle semble ne pas être affectée par les faux négatifs de M tout comme la statistique $z_{M,IBD}$.

4.3.7 Les mesures de comparaison selon les fenêtres

A chaque fenêtre autour d'un marqueur cible on va donc calculer certains indicateurs pour la caractérisation des segments chromosomiques tels que la moyenne du DL, le nombre d'haplotypes recensés et le taux de segments chromosomiques recombinés. A chaque fenêtre on calculera également la moyenne et les écarts types des mesures de comparaison suivantes : $z^{norm}_{M,IBD}, \theta_{M,IBD}$, $\langle IBD - M, M\rangle$ et la $p - value$, et les normes associées aux matrices de similitudes pour une description complète. On comptera aussi les vrais et faux positifs à certains marqueurs afin de mieux rendre compte des mesures de comparaison.

5 Simulations

5.1 Le choix de la fenêtre

On rappelle que le DL est une fonction décroissante du nombre de générations écoulées depuis la génération de base et qu'il admet aussi pour paramètre le taux de recombinaison r, i.e $D_t = e^{-rt}D_0$. Donc si on n'observe pas suffisamment de recombinaisons dans nos simulations, on risque de se retrouver dans une situation où il deviendrait difficile de discriminer entre les s(IBD) dans la capture des statuts IBD par rapport à plusieurs niveaux de DL différents. Dans la pratique les 26 générations du pedigree ne suffisent pas à ce qu'on observe suffisamment de recombinaisons si on garde la densité de marqueurs initialement disponible. De ce fait la fenêtre de 6 SNPs sur les $2m$ chromosomes (ou les $2m$ segments chromosomiques) sur laquelle on applique les s(IBD) a été construite en choisissant un marqueur sur trois le long du chromosome. En d'autres termes on a augmenté la distance génétique d entre les marqueurs et en conséquence on a aussi augmenté le taux de recombinaison r entre les marqueurs.

5.2 Le choix des marqueurs cibles

Avec la description du DL sur les 235 chromosomes de base on s'intéresse aux 5 marqueurs cibles suivants : M_{42}, M_{54}, M_{365}, M_{418}, et M_{492}. Le tableau 5 donne les moyennes de DL, pour un marqueur sur trois à droite et à gauche, calculés avec la mesure du D' à chacun des marqueurs cibles. Il donne également la distance moyenne entre les marqueurs sur la fenêtre autour de chacun des marqueurs cibles.

$Marqueurs$	M_{42}	M_{54}	M_{492}	M_{418}	M_{365}
D'	$0,380$	$0,337$	$0,888$	$0,926$	$0,975$
$moyenne\ distance\ (cM)$	$0,0801$	$0,0526$	$0,0446$	$0,0438$	$0,0294$

Tableau 5 : Le DL aux 5 marqueurs d'intérêts et la distance moyenne sur la fenêtre des 6 marqueurs pour les 235 chromosomes de base

Toutefois quand on obtient des simulations pour une fenêtre de 6 marqueurs autour des marqueurs cibles sur $2m$ chromosomes, les profils du DL résultants autour des marqueurs

cibles sont spécifiques de chaque simulation. A chaque simulation on calcule donc la moyenne du DL pour chacun des marqueurs cibles avec les 6 marqueurs. Ensuite on moyenne ces moyennes pour l'ensemble des simulations réalisées. Les moyennes du DL résultantes, calculées avec la mesure du D', pour 100 simulations, sont données dans le tableau 6 :

$Marqueurs$	M_{42}	M_{54}	M_{492}	M_{418}	M_{365}
D'	$0,445$	$0,449$	$0,883$	$0,925$	$0,961$

Tableau 6 : Le DL aux 5 marqueurs d'intérêts pour les 970 chromosomes fondateurs

Les marqueurs cibles seront utilisés ultérieurement pour simuler la présence d'un QTL pour la cartographie de QTL. L'objectif de ce choix de 5 marqueurs cibles, dans les simulations, est de contraster la capture des statuts IBD par chacune des s(IBD) selon plusieurs cas de profils du DL à gauche et à droite de ces marqueurs. Cela s'inscrit dans l'idée de l'hypothèse de départ : "Le déséquilibre de liaison sera élevé dans une région IBD petite autour du QTL".

5.3 Le choix du nombre de simulations

On a fait le choix de 100 pour le nombre de simulations du pedigree avec "genedropping". Ce choix a été fait pour deux raisons. La première est que 100 simulations suffisent pour la convergence en probabilité de la moyenne empirique des mesures (D', r^2, $\| \ . \ \|$, $z^{norm}_{M,IBD}$, $\theta_{M,IBD}$ etc) et de leurs écarts types. La deuxième est que le temps de calcul pour les 100 simulations est de 4 heures environ ce qui est très raisonnable.
Au marqueur M_{54}, pour la méthode du score seuillé à 0,8 par exemple, les convergences en probabilité de la moyenne empirique de la mesure de comparaison $z^{norm}_{SCORE\_S,IBD}$ et de son écart type sont illustrées par les figures 15 et 16 ci-dessous. Toutes les autres méthodes ont des mesures qui convergent de la même façon à partir de 50 simulations.

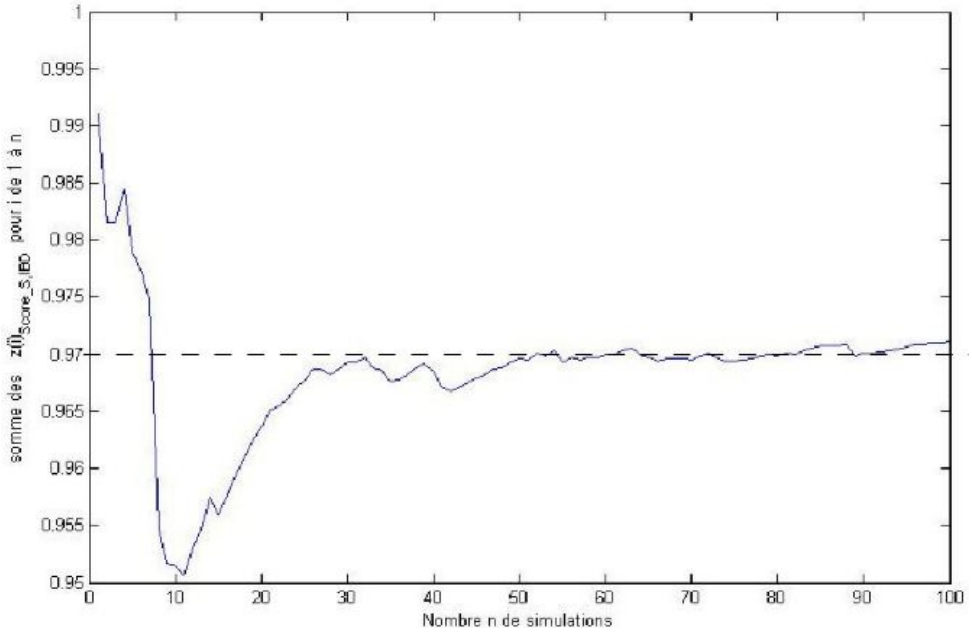

FIGURE 15 – **La convergence pour** $\frac{1}{n}\sum_{i=1}^{n} z^{(i)\ norm}_{SCORE\_S,IBD}$

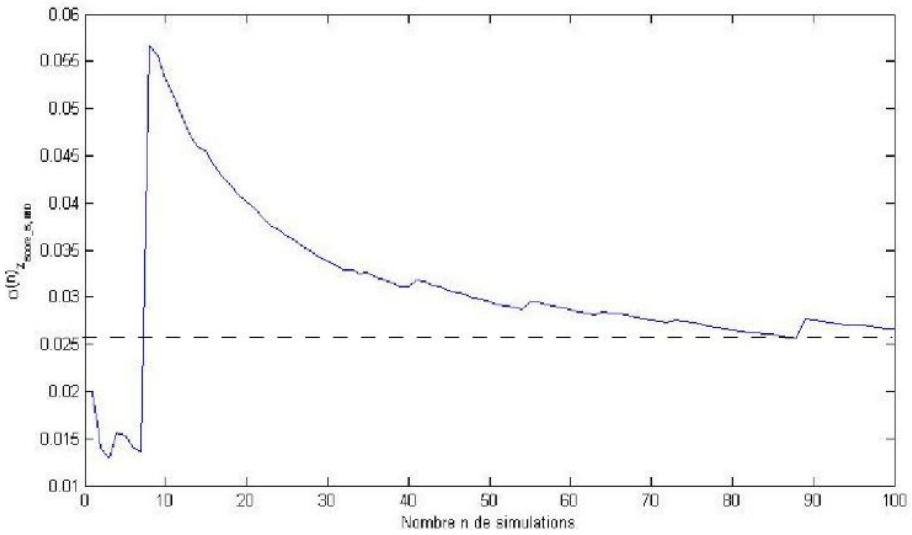

FIGURE 16 – **La convergence pour** $\sigma(n)_{z^{(i)\ norm}_{SCORE\_S,IBD}}$

On remarque que les convergences sont pratiquement atteintes et qu'on a peu de variabilité à partir de 50 simulations pour un écart type inférieur à $0,03$.

5.4 Résultats

Toutes les mesures de comparaisons et les indicateurs pour la caractérisation des segments chromosomiques ont été programmés par moi en Fortran 90 et en MatlabR2009b. La mise en place des simulations a aussi été faite par moi par l'utilisation du Shell Unix. Pour 100 simulations du pedigree avec "genedropping" on obtient les résultats suivants rassemblés dans le tableau 7 et le tableau 8.

Tableau 7 : La norme de M et celle de IBD pour des fenêtres de 6 marqueurs comparées à la position de 5 marqueurs cibles

Marqueurs	M_{42}	M_{54}	M_{492}	M_{418}	M_{365}
Moyenne D'	**0,445**	**0,449**	**0,883**	**0,925**	**0,961**
Moyenne du taux de segments recombinés	**0,110**	**0,074**	**0,067**	**0,059**	**0,040**
Moyenne du nombre de classes d'haplotypes	25	32	13	18	13
Status IBD					
$Moyenne \parallel IBD \parallel_{Euclid=Frob}$	101,456	101,098	101,441	101,303	101,267
$Ecart\ type \parallel IBD \parallel_{Euclid=Frob}$	6,692	6,472	6,203	5,889	6,071
IBS					
$Moyenne \parallel IBS \parallel_{Euclid=Frob}$	229	223	376	309	324
$Ecart\ type \parallel IBS \parallel_{Euclid=Frob}$	16,955	17,010	26,805	20,572	17,533
Score de Similarité					
$Moyenne \parallel SCORE \parallel_{Euclid=Frob}$	409	394	425	428	394
$Ecart\ type \parallel SCORE \parallel_{Euclid=Frob}$	16,867	16,318	19,040	19,627	13,743
Score de Similarité seuillé à 0,8					
$Moyenne \parallel SCORE\_S \parallel_{Euclid=Frob}$	299	259	384	359	332
$Ecart\ type \parallel SCORE\_S \parallel_{Euclid=Frob}$	23,309	14,783	24,924	23,620	18,226
$\mathbb{P}$(IBD)					
$Moyenne \parallel PrIBD \parallel_{Euclid=Frob}$	273	338	442	364	395
$Ecart\ type \parallel PrIBD \parallel_{Euclid=Frob}$	18,100	26,064	31,145	26,198	23,562

Tableau 8 : Les mesures de comparaison de M avec IBD pour des fenêtres de 6 marqueurs comparées à la position de 5 marqueurs cibles

Marqueurs	M_{42}	M_{54}	M_{492}	M_{418}	M_{365}
Moyenne D'	**0,445**	**0,449**	**0,883**	**0,925**	**0,961**
Moyenne du taux de segment recombinés	**0,110**	**0,074**	**0,067**	**0,059**	**0,040**
Moyenne du nombre de classes d'haplotypes	25	32	13	18	13
IBS					
Moyenne $z^{norm}_{IBS,IBD}$	**0,928**	**0,956**	**0,962**	**0,961**	**0,974**
Ecart type $z^{norm}_{IBS,IBD}$	0,040	0,035	0,026	0,026	0,021
Moyenne $\theta_{IBS,IBD}$	**1,145**	**1,121**	**1,307**	**1,249**	**1,261**
Ecart type $\theta_{IBS,IBD}$	0,043	0,040	0,023	0,024	0,024
Moyenne $\langle IBD - IBS, IBS\rangle$	**-43167**	**-40304**	**-132340**	**-85882**	**-95280**
Ecart type $\langle IBD - IBS, IBS\rangle$	7959	7328	20456	12507	11530
$p\text{-}value_{IBS,IBD}$	0,000	0,000	0,000	0,000	0,000
Score de Similarité					
Moyenne $z^{norm}_{SCORE,IBD}$	**0,980**	**0,988**	**0,990**	**0,990**	**0,993**
Ecart type $z^{norm}_{SCORE,IBD}$	0,013	0,009	0,007	0,007	0,017
Moyenne $\theta_{SCORE,IBD}$	**1,325**	**1,314**	**1,332**	**1,334**	**1,313**
Ecart type $\theta_{SCORE,IBD}$	0,018	0,019	0,017	0,015	0,017
Moyenne $\langle IBD - SCORE, SCORE\rangle$	**-157273**	**-145155**	**-171191**	**-173978**	**-145279**
Ecart type $\langle IBD - SCORE, SCORE\rangle$	13861	12763	16457	16743	10769
$p\text{-}value_{SCORE,IBD}$	0,000	0,000	0,000	0,000	0,000
Score de Similarité seuillé à 0,8					
Moyenne $z^{norm}_{SCORE\_S,IBD}$	**0,956**	**0,971**	**0,976**	**0,980**	**0,983**
Ecart type $z^{norm}_{SCORE\_S,IBD}$	0,028	0,027	0,019	0,015	0,016
Moyenne $\theta_{SCORE\_S,IBD}$	**1,24**	**1,181**	**1,309**	**1,290**	**1,266**
Ecart type $\theta_{SCORE\_S,IBD}$	0,032	0,032	0,022	0,023	0,023
Moyenne $\langle IBD - SCORE\_S, SCORE\_S\rangle$	**-80614**	**-57399**	**-138118**	**-119253**	**-100983**
Ecart type $\langle IBD - SCORE\_S, SCORE\_S\rangle$	14315	7487	19466	16770	12269
$p\text{-}value_{SCORE\_S,IBD}$	0,000	0,000	0,000	0,000	0,000
$\mathbb{P}$(IBD)					
Moyenne $z^{norm}_{PrIBD,IBD}$	**0,944**	**0,987**	**0,984**	**0,981**	**0,989**
Ecart type $z^{norm}_{PrIBD,IBD}$	0,031	0,014	0,014	0,018	0,013
Moyenne $\theta_{PrIBD,IBD}$	**1,211**	**1,270**	**1,342**	**1,293**	**1,313**
Ecart type $\theta_{PrIBD,IBD}$	0,032	0,032	0,021	0,027	0,022
Moyenne $\langle IBD - PrIBD, PrIBD\rangle$	**-65167**	**-105225**	**-186015**	**-123353**	**-146170**
Ecart type $\langle IBD - PrIBD, PrIBD\rangle$	9859	17643	27754	19449	18739
$p\text{-}value_{PrIBD,IBD}$	0,000	0,000	0,000	0,000	0,000

Les résultats montrent de manière générale que IBS est la matrice ayant la plus petite norme, que c'est celle qui donne un produit scalaire le moins négatif (le plus proche de 0) selon la caractérisation (3) du théorème de projection, et que c'est celle qui est à moindre angle avec IBD. Toutefois, à un marqueur cible quelconque, on remarque que les matrices $SCORE$, $SCORE\_S$ et $PrIBD$ ont des statistiques $z^{norm}_{M,IBD}$ plus élevées, avec des écart-types plus petits mais qu'elles ont aussi une mesure d'angle plus élevée avec IBD que IBS.

La mesure d'angle plus élevée pour $SCORE\_S$ et $PrIBD$ s'explique par les distributions des éléments des matrices où l'on a d'avantage de faux positifs captés par ces dernières, des 0 de IBD qui sont captés en β avec l'ensemble fini des β possibles inclus dans $[0,8;1]$.

Au marqueur 54 et au marqueur 365, à la simulation 1 par exemple, on obtient les distributions suivantes (tableau 9 et tableau 10) pour la capture des statuts IBD et nonIBD. Les distributions sont données uniquement pour les éléments extra-diagonaux supérieurs car les matrices sont symétriques.

Tableau 9 : Les distributions des éléments des matrices au marqueur 54

Marqueur 54				
Nombre d'éléments de M/IBD	$(2m)^2 = 970 \times 970 = 940900$			
Nombre d'éléments extra-diagonaux de M/IBD	2m(2m-1)/2= 469965 **(100%)**			
Nb de paires de loci IBD	9998 **(2,13%)**			
Nb de paires de loci nonIBD	459967 **(97,87%)**			
Méthode	*Vrais positifs*	*Faux négatifs*	*Faux positifs*	*Vrais négatifs*
IBS	*IBD=1,IBS=1* **2,07%**	*IBD=1,IBS=0* **0,06%**	*IBD=0,IBS=1* **8,49%**	IBD=0,IBS=0 **89,38%**
Score seuillé à 0,6 (=sc6)	*IBD=1*,sc6≥0,6 **2,12%**	*IBD=1*,sc6=0 **$7,66 \times 10^{-3}$%**	*IBD=0*,sc6≥0,6 **40,49%**	*IBD=0*,sc6=0 **57,38%**
Score seuillé à 0,7 (=sc7)	*IBD=1*,sc7≥0,7 **2,12%**	*IBD=1*,sc7=0 **0,01%**	*IBD=0*,sc7≥0,7 **28,88%**	*IBD=0*,sc7=0 **72,82%**
Score seuillé à 0,8 (=sc8)	*IBD=1*,sc8≥0,8 **2,12%**	*IBD=1*,sc8=0 **0,01%**	*IBD=0*,sc8≥0,8 **15,53%**	*IBD=0*,sc8=0 **82,34%**
$\mathbb{P}$(IBD)	*IBD=1,PrIBD=0,99* **2,09%**	*IBD=1,PrIBD=0* **0,03%**	*IBD=0,PrIBD=0,99* **24,85**	*IBD=0,PrIBD=0* **73,09%**

Tableau 10 : Les distributions des éléments des matrices au marqueur 365

Marqueur 365				
Nombre d'éléments de M/IBD	$(2m)^2 = 970 \times 970 = 940900$			
Nombre d'éléments extra-diagonaux deM/IBD	2m(2m-1)/2= 469965 **(100%)**			
Nb de paires de loci IBD	9663 **(2,06%)**			
Nb de paires de loci nonIBD	460302 **(97,94)%)**			
Méthode	*Vrais positifs*	*Faux négatifs*	*Faux positifs*	*Vrais négatifs*
IBS	*IBD=1,IBS=1* **2,03%**	*IBD=1,IBS=0* **0,03%**	*IBD=0,IBS=1* **23,15%**	*IBD=0,IBS=0* **74,80%**
Score seuillé à 0,6 (=sc6)	*IBD=1*,sc6≥0,6 **2,05%**	*IBD=1*,sc6=0 **$8,09 \times 10^{-3}$%**	*IBD=0*,sc6≥0,6 **34,22%**	*IBD=0*,sc6=0 **63,72%**
Score seuillé à 0,7 (=sc7)	*IBD=1*,sc7≥0,7 **2,04%**	*IBD=1*,sc7=0 **0,01%**	*IBD=0*,sc7≥0,7 **25,06%**	*IBD=0*,sc7=0 **72,88%**
Score seuillé à 0,8 (=sc8)	*IBD=1*,sc8≥0,8 **2,04%**	*IBD=1*,sc8=0 **0,01%**	*IBD=0*,sc8≥0,8 **24,28%**	*IBD=0*,sc8=0 **73,67%**
$\mathbb{P}$(IBD)	*IBD=1,PrIBD=1* **2,04%**	*IBD=1,PrIBD=0* **0,02%**	*IBD=0,PrIBD=1* **35,57%**	*IBD=0,PrIBD=0* **62,39%**

On remarque que l'on a une table de contingence 2×2, associée à chacune des méthodes, pour le comptage des états de vérités par rapport à la capture des statuts IBD et nonIBD. En effet on vérifie toujours que :

$$\begin{cases} \textit{Nombre de paire de loci IBD= Vrais positifs + Faux négatifs} \\ \textit{Nombre de paire de loci nonIBD=Faux positifs +Vrais négatifs} \end{cases}$$

Les seuls éléments de la matrice M qui ont pour effet de rendre cette dernière dissemblable à la matrice IBD sont ceux ayant un état de vérité faux par rapport aux éléments d'IBD, à savoir les faux négatifs et les faux positifs. Il est donc immédiat de constater que ce sont les faux positifs qui l'emportent face aux faux négatifs car ils sont très largement majoritaires. Ce sont donc ces éléments qui expliquent une mesure d'angle élevée par rapport à la matrice IBD ou un produit scalaire très négatif selon la caractérisation

(3) du théorème de projection. On remarquera aussi que l'on ne s'intéresse qu'aux vrais et faux positifs car par dualité on a que :

$$\begin{cases} \textit{Augmenter les vrais positifs} \Longleftrightarrow \textit{Diminuer les faux négatifs} \\ \textit{Diminuer les faux positifs} \Longleftrightarrow \textit{Augmenter les vrais négatifs} \end{cases}$$

Comme les vrais positifs sont assez bien détectés il vient que le problème du meilleur estimateur de la matrice IBD par les matrices de similarité, associées aux méthodes pour cette étude, est un problème de minimisation du nombre de faux positifs.

La figure 17 ci-dessous montre une configuration que l'on retrouve très souvent pour les matrices de similarité par rapport à IBD. On remarque que la matrice $SCORE$ donne un plus grand produit scalaire avec IBD que les autres matrices. Toutefois cette configuration n'est pas toujours ainsi selon la fenêtre où l'on se place et le niveau du DL associé à cette fenêtre. Il peut arriver par exemple qu'on puisse inverser les rôles de $SCORE$ et de $PrIBD$ par rapport à leurs normes, mais sans que cette dernière donne pour autant un plus grand produit scalaire avec IBD car la matrice $PrIBD$ serait alors à un angle plus élevé avec IBD, (le marqueur M_{492} illustre cette situation dans le tableau 8). De plus on remarque que dans la plupart des cas la matrice du score seuillé à 0,8 ($SCORE\_S$) est de plus petite norme que $PrIBD$, qu'elle est à moindre angle avec IBD que $PrIBD$ et que son produit scalaire avec IBD est comparable à celui de $PrIBD$ avec IBD pour un niveau de DL très élevé (les marqueurs M_{365} et M_{418} illustre cette situation dans le tableau 8).

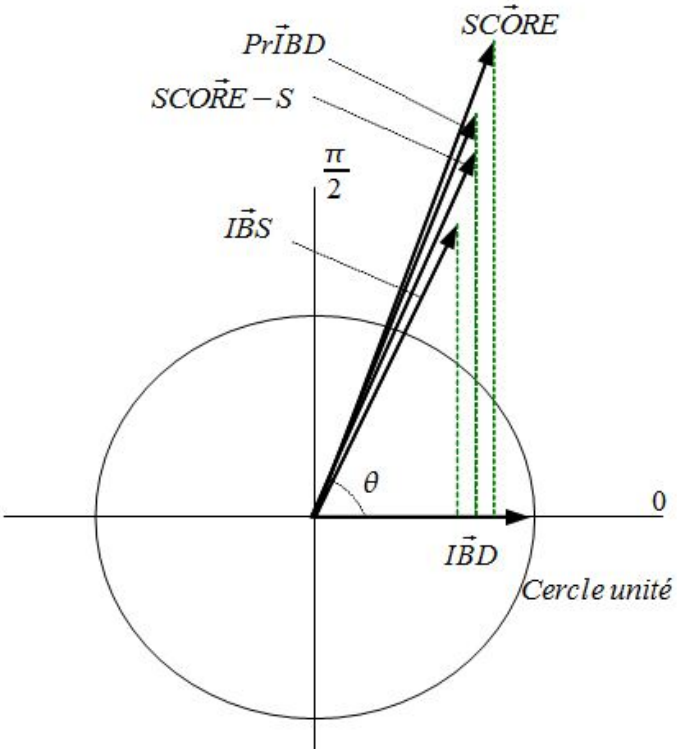

FIGURE 17 – **Exemple de projection et d'angle de chacune des matrices par rapport à** IBD

On constate aussi de manière générale que $SCORE\_S$ et $PrIBD$ détectent mieux les vrais positifs lorsque le niveau du DL est plus élevé autour d'un marqueur. Toutefois pour un

niveau de DL plus élevé elles détectent aussi plus de faux positifs car on s'éloigne d'IBD avec une mesure d'angle plus élevé. Cette tendance est illustrée pour ces méthodes en comparant leurs mesures de comparaisons par rapport à l'IBD au marqueur 54 (tableau 8), où l'on a un niveau de DL moyen, avec celles du marqueur 365 (tableau 8) où l'on a un niveau de DL élevé par exemple. De plus on voit aussi que les mesures d'angles les plus basses sont atteintes lorsque le niveau du DL est au plus bas pour les 5 marqueurs d'intérêts.

Au marqueur 365 on constate également que la matrice $SCORE\_S$ est dans une situation de presque homothétie avec IBS ($\theta_{SCORE\_S,IBD} = 1,266 \approx 1,261 = \theta_{IBS,IBD}$) et qu'elle détecte plus de vrai positifs que IBS ($z^{norm}_{SCORE\_S,IBD} = 0,983$ et $z^{norm}_{IBS,IBD} = 0,973$). Cette situation est celle décrite au chapitre 4 précédent (sous-section 4.2.4) pour $M_1 = SCORE\_S$ et $M_2 = IBS$, elle est donc un meilleur estimateur d'IBD que IBS dans ce cas précis avec un niveau de détection des vrais positifs à $0,983$ et un ratio de vrais/faux positifs à $0,084$ (obtenu du tableau 10) comparable à ce même ratio pour l'IBS qui est de $0,087$ avec un niveau de détection à $0,973$. La situation de presque homothétie de $SCORE\_S$ avec IBS s'explique en comparant la distribution des éléments de chacune des 2 matrices où l'on a une compensation entre les rapports relatifs des vrais et faux positifs. En effet les faux positifs de $SCORE\_S$ correspondent à des valeurs de $s_{i,j}$ appartenant à un ensemble inclus dans $[0,8; 1]$ tandis que les faux positifs de IBS sont tous détectés à $s_{i,j} = 1$. La plupart des vrais positifs détectés par $SCORE\_S$ sont à 1 au lieu de $0,8$. En effet si deux haplotypes sont IBS alors le score de similarité est maximal et égal à 1 (après normalisation), ce qui est égal à l'IBS.

De manière générale on voit aussi que $SCORE\_S$ est la matrice qui est la plus proche de IBS par rapport à sa mesure d'angle avec IBD, sauf pour le niveau de DL le plus bas à M_{42}, et qu'elle détecte toujours plus de vrais positifs que IBS pour un marqueur quelconque. $PrIBD$ est quant à elle la matrice la plus éloignée de IBS par rapport à la caractérisation (3) ou la mesure d'angle. Toutefois à M_{42} elle est plus proche de IBS pour ces mesures que les autres matrices mais elle détecte tout de même moins de vrais positifs que $SCORE\_S$ à ce marqueur. La matrice $PrIBD$, étant déjà constituée d'attributs binaires (0 ou 1, 0 ou 0,97 par exemple), ne peut pas être seuillée et semble la moins efficace pour l'estimation de la matrice IBD.

6 Discussion

6.1 La provenance des faux positifs

Le fait d'avoir d'avantage de faux positifs pour un DL plus élevé à une fenêtre s'explique par le fait qu'on ait des segments chromosomiques qui se ressemblent et qui sont d'origines chromosomiques différentes. En effet, on remarque par exemple que le DL au marqueur 365 était initialement égal à 0,975 pour les 235 chromosomes de base, et que les 235 segments chromosomiques en ségrégation sur la fenêtre associée à ce marqueur se ressemblaient déjà avant d'effectuer les "genedropping" à chaque simulation (13 haplotypes en ségrégation autour du marqueur 365 dans tableau 8). On a donc, à chaque "genedropping" le long du pedigree pour un nombre fini de générations (26), une forte probabilité d'obtenir des segments chromosomiques en ségrégation sur une fenêtre qui se ressemblent et qui sont d'origines chromosomiques différentes, dès lors que le niveau du

DL est élevé à cette fenêtre.

Lorsque l'on a un faible niveau de DL à une fenêtre on obtient dans de la même façon un peu plus de segments chromosomiques qui ne se ressemblent pas et qui sont d'origines chromosomiques différentes. On peut ainsi mieux distinguer les $2m$ segments chromosomiques par rapport à leur chromosome d'origine dès lors que certains se ressemblent, ce qui crée moins de faux positifs. Toutefois on remarquera que cette distinction ne se fait que dans une certaine mesure. En effet selon la méthode utilisée (l'IBS, score de similarité seuillé à 0,8 et $\mathbb{P}$(IBD) par exemple), on arrive à détecter moins de faux positifs lorsque le niveau de DL est plus faible .

6.2 Une approche de minimisation des faux positifs

Le nombre important de faux positifs est un problème majeur pour cette étude. Les 26 générations du pedigree utilisées lors des simulations ne suffisent certainement pas à particulariser les segments chromosomiques afin de leur donner une signature par rapport aux évènements de recombinaisons, pour qu'ils soient d'avantage repérable comme étant au moins porteurs d'un allèle IBD et en l'occurrence un allèle IBD au QTL. En effet deux segments chromosomiques recombinés qui se ressemblent à l'issue d'un "genedropping" ont plus de chances d'avoir la même trace dans le pedigree et donc de partager un allèle IBD d'un chromosome fondateur. En ce sens augmenter la taille du pedigree serait éventuellement une bonne chose mais on diminuerait aussi le DL aux marqueurs cibles étudiés, et on ne serait alors plus dans l'hypothèse de départ. En outre rien ne garantit aussi qu'on diminuerait substantiellement le nombre de faux positifs aux endroits où l'on observe un fort DL si on augmente la taille du pedigree. En effet les segments chromosomiques se ressemblent aux endroits où l'on observe un fort DL et les évènements de recombinaisons ne risquent pas de créer substantiellement de nouveaux haplotypes. De plus si on considère que l'on a $2^6 = 64$ haplotypes possibles pour 6 marqueurs à une fenêtre, proportionnellement aux 970 segments chromosomiques en ségrégation sur cette fenêtre, alors on peut considérer qu'il existera encore un risque non négligeable pour la détection des faux positifs.

Il vient donc que faire varier avec modération le nombre de marqueurs (7 à 14 marqueurs) constituant la fenêtre sur les $2m = 970$ chromosomes fondateurs, en considérant le DL calculé localement pour une sous-fenêtre incluse au centre de cette fenêtre (14 marqueurs par exemple), jusqu'à ce que l'on obtienne les meilleurs résultats possibles, pourrait s'avérer une bonne stratégie. On devrait également choisir un ensemble de chromosomes de base où les chromosomes seraient les plus différents possibles, car la race porcine Large White est issue de l'équivalent génétique d'un petit nombre d'individus (≈ 50) non apparentés. Une autre stratégie afin de mieux répartir les statuts IBD entre les 970 segments chromosomiques serait de constituer un ensemble bien plus grand de chromosomes de base, par des tirages aléatoires, à partir de des 235 chromosomes existants. En ce sens on diminuerait éventuellement le fait d'avoir des segments chromosomiques qui se ressemblent en étant d'origine chromosomique différente.

6.3 Les développements et limites possibles des méthodes

Des méthodes utilisées pour quantifier la ressemblance entre les segments chromosomiques, l'IBS semble la plus efficace pour se rapprocher du vrai statut IBD entre ces derniers. Toutefois cette modélisation est limitée par son aspect binaire (fonction booléenne) et son potentiel de prédiction est donc déjà à son maximum. Au vu des résultats le calcul de $\mathbb{P}$(IBD) quant à lui semble faire des hypothèses selon un modèle de coalescence qui n'ont pas l'air d'être réellement respectées. Pour des relations de non apparentés chez les fondateurs pour la cartographie de QTL les probabilités calculées varient dans un ensemble binaire et on a l'intuition, par rapport aux matrices $PrIBD$ obtenues dans la pratique, que les calculs se font essentiellement sur la base de l'information IBS des marqueurs adjacents. Comme pour l'IBS, $\mathbb{P}$(IBD) semble déjà être à son potentiel de prédiction maximal car d'une part il est à valeurs dans un ensemble binaire et d'autre part les hypothèses du modèle de coalescence sur lesquelles repose ce calcul ne semblent pas être réellement respectées.

Le score de similarité quant à lui semble ne pas avoir atteint les limites de son potentiel de prédiction. En effet les résultats montrent qu'on peut construire à partir de cette dernière une modélisation, le score de similarité seuillé à 0,8, qui respecte d'avantage l'hypothèse de départ à savoir que "Le déséquilibre de liaison sera élevé dans une région IBD toute petite autour du QTL". Toutefois la structure des données ne permet pas de le montrer d'une manière très évidente à cause des faux positifs, d'où la nécessité d'avoir des simulations aussi réalistes que possible et de faire les bonnes hypothèses par rapport à celles-ci. On pourrait donc adopter d'autres formulations du score de similarité, où les fonctions poids w_1 et w_2 seraient des fonctions exponentielles avec un paramètre α ($\alpha \in \mathbb{R}^{-*}$) par exemple qu'on pourrait faire varier sur un ensemble de valeurs à définir, afin d'obtenir une décroissance plus ou moins rapide de la pondération des marqueurs voisins lorqu'on s'éloigne du locus de référence (le processus de lissage exponentiel).

Une autre voie envisageable est le calcul de la fonction de krigeage dans le cadre d'une prédiction gaussienne ou d'un champs gaussien. Dans le cadre d'une prédiction gaussienne cette fonction calcule l'espérance d'une variable aléatoire gaussienne Y (centrée) conditionnellement à un vecteur gaussien X d'observations, en utilisant la matrice de variance-covariance du vecteur X, Γ_X, et le vecteur de covariance de Y selon chacune des composante de X, $U_{X,Y}$. La fonction de krigeage est définie de la façon suivante $\mathbb{E}(Y|X) = U_{X,Y}^T \Gamma_X^{-1} X$. Une adaptation de cette fonction peut donc être envisagée. En effet on pourrait calculer l'espérance du statut IBD à un locus conditionnellement à l'haplotype recensé et utiliser un seuil de décision pour affecter le statut IBD ou non. Cette voie reste à explorer.

6.4 Une seule mesure de comparaison.. ?

La comparaison des matrices de similarité avec la matrice IBD a nécessité un ensemble de mesures qui ont montré leurs complémentarités. Toutefois pour tester la minimisation de l'écart entre une matrice $M \in E_s$ avec IBD sur un grand nombre de paramètres, pour le lissage exponentiel du score de similarité pour $\alpha \in [\beta_1; \beta_2]$ par exemple, il est à se demander s'il n'est pas impossible de construire une seule mesure (une combinaison éventuelle) qui puisse évaluer à elle seule cet écart. Le ratio de vrais/faux positifs (ou

mesure de prédiction) qu'on appelle $\rho_{M,IBD}$ n'est pas satisfaisant car pour un très petit nombre de faux positifs et pour un petit nombre de vrais positifs (mais suffisamment grand) ce rapport devient grand et il ne traduit pas la situation réelle de ressemblance entre M et IBD. Une mesure qui permettrait de pallier à ce problème est la suivante :

$$Z = \mathbb{1}_{\{z_{M,IBD}>\alpha\}}\rho_{M,IBD};\ \ Z \in]0;+\infty[$$

$$\text{avec } \rho_{M,IBD} = \frac{Nb.VP_M}{Nb.FP_M} \text{ et } Nb.FP_M = \begin{cases} Nb.FP_M \text{ si } Nb.FP_M \neq 0 \\ \frac{1}{2} \text{ si } Nb.FP_M = 0 \end{cases}$$

$$\text{où : } \begin{cases} Nb.VP_M = \text{ nombre de vrais positifs de } M \\ Nb.FP_M = \text{nombre de faux positifs de } M \\ \alpha = \text{le \% minimum de vrais positifs pour } M \text{ que l'on souhaite obtenir} \end{cases}$$

Il vient alors que l'estimateur naturel d'IBD pour un seuil α du pourcentage de vrais positifs est donné par :

$$\hat{IBD} = \underset{M \in E_s}{argmax} \left(\mathbb{1}_{\{z_{M,IBD}>\alpha\}}\rho_{M,IBD}\right)$$

Une autre mesure qui permettrait de minimiser le nombre de faux positifs et de maximiser le nombre de vrais positifs (par rapport à un seuil défini) peut être définie de la façon suivante :

$$Z = \mathbb{1}_{\{Nb.VP_M>\beta\}} \left(\frac{Nb.VP_M + (Nb.nonIBD_M - Nb.FP_M)}{Nb.Total_M} \right);\ \ Z \in [0;1]$$

$$\text{où : } \begin{cases} Nb.nonIBD_M = \text{ nombre de paires de loci nonIBD de } M \\ Nb.Total_M = \text{ nombre total d'éléments de } M \\ \beta = \text{le nombre minimum de vrais positifs pour } M \text{ que l'on souhaite obtenir} \end{cases}$$

Pour Z proche de 1 le nombre d'états de vérités vrais est presque maximal et inversement pour le nombre d'états de vérités faux. Il vient alors que l'estimateur naturel d'IBD pour un seuil α du nombre de vrais positifs est donné par :

$$\hat{IBD} = \underset{M \in E_s}{argmax} \left(\mathbb{1}_{\{Nb.VP_M>\beta\}} \left(\frac{Nb.VP_M + (Nb.nonIBD_M - Nb.FP_M)}{Nb.Total_M} \right) \right)$$

Ces estimateurs viennent du fait que plus de 90% des vrais positifs semble être détectés pour toutes les régions du génome et que le problème de trouver un bon estimateur d'IBD est un problème de minimisation du nombre de faux positifs. La mesure Z quant à elle nous permet surtout de regarder l'écart entre M et IBD pour un ensemble de paramètres comme décrit précédemment.

7 Conclusion

L'étude a mis en évidence que plus de 95% des segments chromosomiques ayant un allèle IBD au QTL sont détectés par l'IBS, le score de similarité seuillé à 0,8 et $\mathbb{P}$(IBD) lorsque l'on se trouve dans une région (fenêtre) où il existe un fort DL. Si on ne considère pas l'IBS parmi ces méthodes de similitude alors plus de 98% des segments chromosomiques ayant un allèle IBD au QTL sont détectés dans une région où il existe un fort DL. Toutefois l'étude a aussi mis en évidence le problème des faux positifs détectés par ces méthodes qui sont bien plus nombreux que les vrais positifs pour tout niveau de DL à une fenêtre quelconque. Le problème de trouver un bon estimateur de la matrice IBD par les matrices associées à ces méthodes pour cette étude est un problème de minimisation des faux positifs comme évoqué dans la partie Résultats. De toute les méthodes étudiées l'IBS est celle qui détecte le moins de faux positifs. Toutefois la méthode du score seuillé à 0,8 s'avère meilleure que l'IBS dans une région où il existe un fort DL. De ce fait on a trouvé une méthode qui donne un meilleur estimateur (relative aux autres méthodes étudiées), la matrice $SCORE\_S$, de la matrice IBD pour une région où il existe un fort DL et qui tient d'avantage compte de l'hypothèse de départ : "Le déséquilibre de liaison sera élevé dans une région IBD toute petite autour du QTL". Néanmoins, comme les résultats l'ont montré, les matrices produites par les s(IBD) étudiées à une fenêtre (E_s) sont encore très dissimilaires à la matrice IBD. Il reste donc encore des développements et stratégies à définir, et à expérimenter, comme celles de 6.2, 6.3 et 6.4 de la partie Discussion afin de trouver un bon estimateur de la matrice IBD. L'estimation de la matrice IBD ainsi obtenue à chaque fenêtre sur le génome pourra être utilisée en cartographie de QTL où l'on testera sa performance.

8 Bibliographie

Boitard, S. (2006). Cartographie de gènes à caractères quantitatifs par déséquilibre de liaison. *Thèse doctorale en mathématiques appliquées, option statistique, de l'Université Toulouse III.*

Bidanel J.-P., Boichard D., Chevalet C. (2008). De la génétique à la génomique. *INRA Production animale*, 21(1), 15-32.

Cierco-Ayrolles, C., Abdallah, J., Boitard, S., Chikhi, L., de Rochambeau, H., Tsitrone, A., Veyrieras, J., and Mangin, B. (2004). On linkage disequilibrium measures : methods and applications in recent research developments in genetics and breeding. *Research SignPost, Kerala, India*, **1** : 151-180.

Escoufier Y., L'Hermier H. (1978). A Propos de la Comparaison Graphiques des Matrices de Variance. *Biom.J.*, Vol.20 no.5, 477-483.

Heifetz E.M., Fulton J.E., O'Sullivan N., Zhao H., Dekkers J.C.M., Soller M., 2005. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics*, 171 : 1173-1181.

Hill, W.G., Robertson A. (1968) Linkage disequilibrium in finite populations. Theor. Appl. *Genetics*, 38, 226-231.

Lewontin R.C. (1964) The interaction of selection and linkage. I. *Genetics*, 49, 49-67.

Lewontin R.C., Kojima K. (1960) The evolutionary dinamics of complex polymorphisms. *Evolution*, 14, 458-472.

Li J., Jiang T. (2005). Haplotype-based linkage desequilibrium mapping via direct data mining. *Bioinformatics*, Vol.21 no.24, 4384-4393.

Li J., Zhou Y., Elston R.C. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm. *BMC Bioinformatics*, **7** :258

Meuwissen T.H.E, Goddard M.E. (2007). Multipoint Identity-by-Descent Prediction Using Dense Markers to Map Quantitative Trait Loci and Estimate Population Size. *Genetics*, **176** : 2551-2560 .

Meuwissen T.H.E, Goddard M.E. (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.* Vol.33, 605-634 .

Scheet P., Stephens M. (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data : Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics*, Vol.78.

Schneider J.W., Borlund P. (2007). Matrix Comparison, Part1 : Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology.*

Schneider J.W., Borlund P. (2008). Matrix Comparison, Part2 : Measuring the resemblance between proximity measures or ordination results by the use of Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology.*

Ytournel F. (2008). Déséquilibre de liaison et cartographie de QTL en population sélectionnée. *Thèse doctorale en génétique animale d'AgroParisTech.*

Ytournel F., Gilbert H., Boichard D. (2008). Comment affiner la localisation d'un QTL ? *INRA Production animale*, 21(2), 147-158.