UNIVERSITÉ DU QUÉBEC À MONTRÉAL

COMPARAISON ENTRE UNE MÉTHODE DE CARTOGRAPHIE FINE ET DES TESTS D'ASSOCIATION DANS LE CADRE DE L'ÉTUDE D'UN CARACTÈRE GÉNÉTIQUE AVEC DES DONNÉES INCOMPLÈTES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

NA YANG

JUILLET 2019

REMERCIEMENTS

Je tiens d'abord à remercier Fabrice Larribe, mon directeur de recherche. Merci pour son soutien, sa créativité inspirante et ses idées innovatrices. J'aimerais également remercier Sorana Froda, ma codirectrice, pour sa générosité, sa patience et ses encouragements constants.

Je voudrais de plus remercier tous les membres de l'ÉMOSTA. Merci pour la passion qu'ils mettent dans leur travail.

Je tiens aussi à remercier le corps de soutien technique et administratif du Département de mathématiques de l'UQAM. Merci pour son support durant toutes mes études.

Finalement, un merci à toute ma famille et tous mes amis. Merci pour votre présence et votre soutien.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Nous présentons et testons deux méthodes permettant d'estimer la position d'une mutation causale le long d'une séquence génétique dans un échantillon en présence de données manquantes. La première méthode, DMap, s'insère dans la méthodologie de la cartographie génétique fine, basée sur le processus de coalescence. Cette méthode est capable d'analyser des données incomplètes directement en les considérant non-informatives. La deuxième méthode est un test d'association qui utilise des données imputées par la méthode IMPUTE. Ce type de test (SNPTEST) est beaucoup appliqué dans les études d'association sur le génome. On compare ces deux méthodes par une étude de simulation selon plusieurs scénarios. On constate que la méthode DMap a une performance comparable à la méthode SNPTEST. Si environ 50% de données sont manquantes dans l'échantillon, il sera plus avantageux d'utiliser la méthode DMap.

MOTS-CLÉS : SNP, cartographie génétique, processus de coalescence, ARG, fonction de pénétrance, imputation, test d'association.

INTRODUCTION

Ce mémoire porte sur des études d'association entre des variations sur le génome et un trait d'intérêt (maladie, par exemple). Dans une telle étude génétique, il arrive souvent que plusieurs génotypes de certains individus soient manquants dans un échantillon en raison d'un défaut expérimental. Pour pouvoir trouver le lien entre des variations génétiques et ce caractère d'intérêt avec ce genre de données, on choisit en général une méthode d'imputation permettant d'inférer les génotypes manquants afin d'appliquer un test d'association. Cette méthode utilise un panel de référence qui contient des génotypes provenant de quelques individus sélectionnés dans une population type, génotypes qui se situent dans la même région du génome que les génotypes manquants. Malgré que le test d'association (SNPTEST) analyse des données imputées en prenant en considération l'incertitude de l'imputation, l'information obtenue à travers le panel de référence est limitée et cela peut causer une imputation inexacte, ce qui a un impact sur le test d'association.

Une réponse à ce type de problème est d'utiliser une méthode qui ne demande pas de compléter les données manquantes afin de faire les analyses : DMap. C'est une méthode de cartographie génétique fine qui s'appuie sur la théorie de la coalescence et construit des généalogies des individus de l'échantillon sans nécessairement connaître tous les génotypes. L'objectif de ce mémoire est de décrire et d'évaluer ces deux méthodes qui servent à analyser les données génétiques incomplètes dans un échantillon. Nous voulons surtout comparer la performance de la méthode DMap avec celle de la méthode SNPTEST en testant leur sensibilité à plusieurs

paramètres, dont la fréquence de la maladie dans la population et la proportion de données manquantes dans l'échantillon.

Ce mémoire est divisé en 6 chapitres. Le premier chapitre contient une introduction à quelques concepts de base en génétique qui seront utilisés dans la suite. Le deuxième chapitre présente la théorie de la coalescence, tandis que la méthode DMap est présentée en détails au troisième chapitre. Au quatrième chapitre, nous décrivons la méthode IMPUTE qui permet d'inférer les génotypes manquants dans un échantillon. Par la suite, la méthode SNPTEST est présentée au cinquième chapitre. Nous terminons ce mémoire, avec le sixième chapitre, qui présente une étude de simulation afin de comparer les résultats obtenus par les deux méthodes, DMap et SNPTEST.

CHAPITRE I

NOTIONS DE GÉNÉTIQUE DES POPULATIONS

Le terme génétique a été inventé par le biologiste anglais William Bateson (1861-1926) en 1905. L'objectif était de décrire la science de l'hérédité et de la variation d'une espèce dans un mot. Ce chapitre présente des termes et concepts basiques en génétique chez l'humain qui sont utilisés dans ce mémoire.

1.1 Définition des termes de base en génétique chez l'humain

1.1.1 Le chromosome

Le corps humain est composé de milliards de cellules, et le noyau de chaque cellule contient des chromosomes. Les cellules humaines peuvent contenir soit 46 chromosomes, soit 23 chromosomes. Les *cellules somatiques* (non sexuelles) possèdent 46 chromosomes, sous la forme de 23 paires de chromosomes. Ce type de cellule est dite *diploïde* car il a deux cellules parentales. Les *gamètes* (cellules reproductives) possèdent un exemplaire de chacune de ces 23 paires de chromosomes, et ils sont dits *haploïdes*. Les 22 premières paires sont les autosomes et la $23^{\text{ème}}$ paire est composée des chromosomes sexuels. Les chromosomes sexuels sont différents par rapport au sexe : deux chromosomes identiques chez les femmes et deux différents chez les hommes. Les chromosomes d'une même paire sont dits aussi *homologues*.

1.1.2 L'ADN

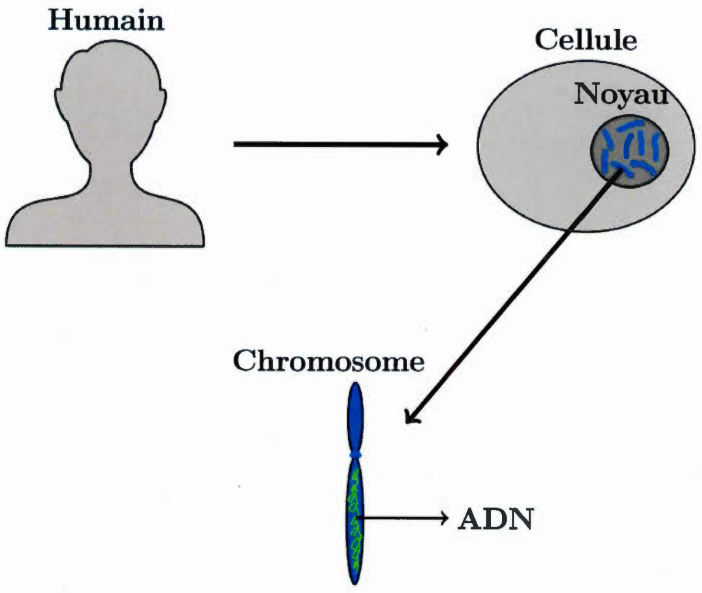


Figure 1.1: Illustration schématique d'un chromosome et sa relation avec l'ADN.

L'information génétique contenue dans les chromosomes est représentée par l'*ADN* (voir Figure 1.1). La molécule d'ADN, ou l'acide désoxyribonucléique, est une longue molécule qui est constituée de deux chaînes de *nucléotides* sous une forme de double hélice. Chaque nucléotide est constitué d'un acide phosphorique, d'un sucre (le désoxyribose) et d'une base azotée. Il existe 4 différentes bases azotées possibles :

— l'adénine que l'on peut noter A ;

— la thymine que l'on peut noter T ;

— la cytosine que l'on peut noter C ;

— la guanine que l'on peut noter G.

Ces bases azotées s'attachent en respectant une règle de complémentarité : l'adénine et la thymine sont complémentaires, la cytosine et la guanine sont

complémentaires. Donc, A est toujours reliée avec T, et C est toujours reliée avec G (voir Figure 1.2). Ainsi, si l'on sépare les deux chaînes d'ADN, l'une peut être utilisée pour reconstituer l'autre. Ceci contribue à la réplication de l'information génétique.

1.1.3 Les gènes

Un *gène* est un segment d'ADN codant. C'est donc un petit fragment d'un chromosome, et il occupe une position précise sur le chromosome, cette position est appelée *locus* (loci au pluriel). L'ensemble du matériel génétique d'un organisme est appelé *génome*, c'est-à-dire l'ensemble des régions codantes (gènes) et des régions non-codantes de cet organisme.

La longueur des molécules d'ADN est mesurée par le nombre d'appariements de deux bases azotées (ou deux nucléotides) sur les deux chaînes complémentaires d'ADN, cela se nomme aussi *distance physique*. Une paire de bases azotées est notée par pb, en anglais «*base pair*». Par exemple, la longueur de la séquence d'ADN dans la figure 1.2 est de 9 pb. Le génome humain contient environ 3.2 milliards de paires de nucléotides. De ce fait, on utilise plutôt la *mégabase* (Mb) pour mesurer la longueur d'ADN, elle correspond à 1 million de paires de bases.

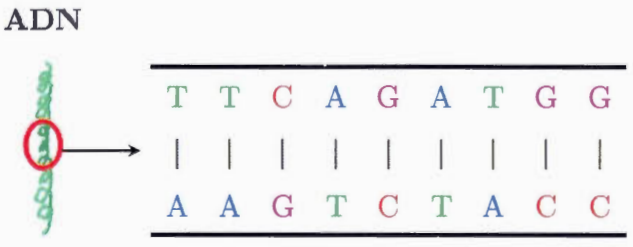


Figure 1.2: Illustration schématique d'une séquence d'ADN.

Un gène peut posséder plusieurs versions, et chaque version est appelée *allèle*.

Chaque humain possède deux allèles sur le même locus, un allèle vient du père et l'autre vient de la mère. La composition des allèles d'une personne est appelée *génotype*, celui-ci peut influencer un caractère détectable d'un humain, c'est-à-dire, le *phénotype*. Un génotype peut être composé de deux allèles identiques ou de deux allèles différents. Quand c'est le deuxième cas, le phénotype correspond à l'effet d'un allèle dit *dominant*, et l'autre allèle est dit *récessif*. La figure 1.3 donne un exemple sur le lien entre un génotype et un phénotype. Un gène qui se situe sur la 22ème paire de chromosomes détermine la présence ou l'absence du lobe libre de l'oreille. Notons par A l'allèle dominant qui détermine la présence du lobe libre et a l'allèle récessif qui détermine son absence (lobe attaché). Ainsi, les phénotypes qui correspondent aux 3 génotypes AA, Aa et aa sont respectivement : lobe libre, lobe libre et lobe attaché.

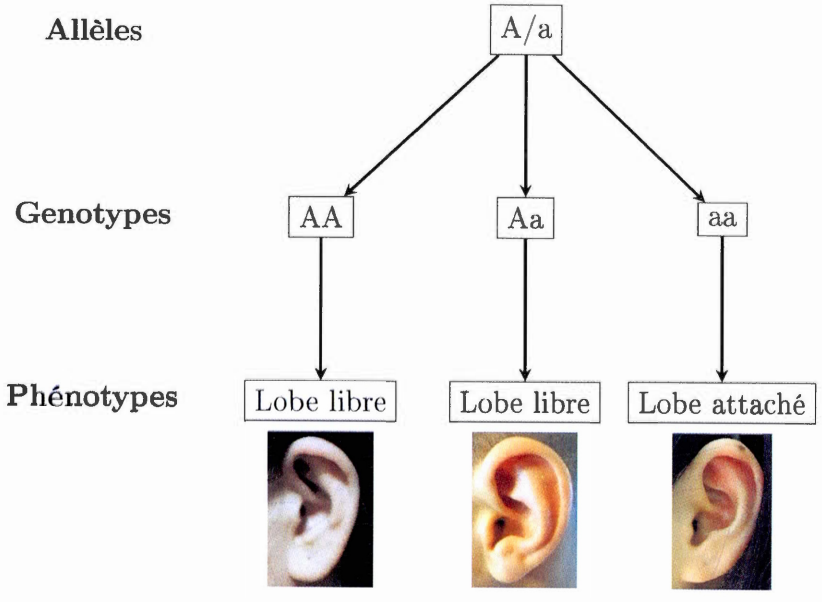


Figure 1.3: Exemple du mode d'attache du lobe de l'oreille selon les différents génotypes. Les photos sont prises par l'auteur du mémoire.

1.2 Hérédité et variabilité génétique chez l'humain

En général, dans une famille avec plusieurs enfants, les enfants se ressemblent tous à un certain degré, tout en ressemblant à leurs parents. Toutefois, chaque enfant possède ses propres particularités. Ces deux aspects contradictoires de la reproduction biologique peuvent être expliqués par l'hérédité et la variabilité génétique.

1.2.1 Hérédité

L'hérédité est la transmission des caractères d'une génération à la suivante au sein d'une espèce. Chez l'humain, les spermatozoïdes et ovules sont les gamètes respectifs de l'homme et de la femme. Un spermatozoïde et un ovule fusionnent pendant la fécondation, et ils forment une seule cellule qui s'appelle *zygote*. Ainsi, un zygote hérite des informations génétiques de ses deux parents. Dans chaque paire de chromosomes homologues de cette cellule, une copie vient de la mère et l'autre vient du père. De ce fait, les deux copies d'une paire de chromosomes ne sont pas forcément identiques, car ils n'ont pas les mêmes gènes à chaque locus. La combinaison de deux gènes identiques sur un locus est dite *homozygote*, et la combinaison de deux gènes différents sur un locus est dite *hétérozygote*. Un zygote va se développer en un individu en se multipliant constamment, et ce processus est dit division cellulaire. Chez l'humain, il existe deux types de division cellulaire : *mitose* et *méiose*.

La mitose s'applique aux cellules somatiques. Pendant la mitose, une cellule mère se divise en deux cellules filles identiques qui sont aussi identiques à la cellule mère. Donc en général, il n'y a aucun changement d'information génétique par ce type de division cellulaire. C'est la méiose où se produit la variabilité génétique.

1.2.2 Variabilité génétique

La méiose est un type de division cellulaire qui ne produit que des gamètes. Par la méiose, une cellule mère dite *cellule germinale* se divise en quatre cellules filles (gamètes) différentes qui sont aussi différentes de la cellule mère. La méiose contient deux étapes de divisions successives, et la variabilité génétique se crée avant et pendant la première division.

Au début de la méiose, la cellule germinale réplique son ADN, c'est-à-dire chaque chromosome se dédouble en deux *chromatides* complètement identiques. Ensuite, juste avant la première division, des chromosomes homologues dédoublés d'une même paire peuvent s'échanger une partie de leur ADN. Ce comportement s'appelle *enjambement*, et il crée une grande variation génétique. La figure 1.4 illustre un événement d'enjambement entre une paire de chromosomes homologues juste avant la première division de la méiose.

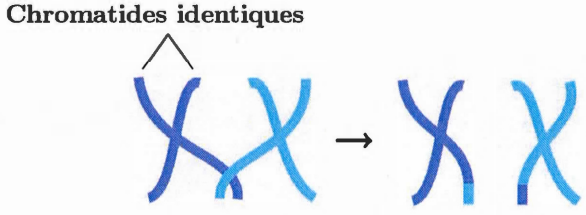


Figure 1.4: Illustration schématique d'un enjambement entre deux chromosomes homologues juste avant la première division de la méiose. Ici, chaque chromosome est constitué de deux chromatides identiques.

Après l'enjambement, une autre variation génétique a lieu. Lors de la première division de la méiose, les chromosomes homologues dédoublés dans une cellule germinale se séparent l'un de l'autre, et ils vont dans deux cellules filles différentes. Ce processus se fait au hasard, il est donc impossible de connaître pour une paire de

chromosomes homologues doubles, quel chromosome va dans une cellule et lequel dans l'autre. Comme il y a 23 paires de chromosomes dans chaque cellule humaine, le nombre de combinaisons possibles des chromosomes dans une cellule fille est de 2^{23}. Cela veut dire qu'il existe plus de 8 millions de gamètes possibles chez une paire de parents. Ceci contribue donc beaucoup à la variabilité génétique. La figure 1.5 nous montre un exemple de 4 combinaisons possibles des chromosomes dans une cellule fille après la première division de la méiose d'une cellule mère de deux paires de chromosomes homologues.

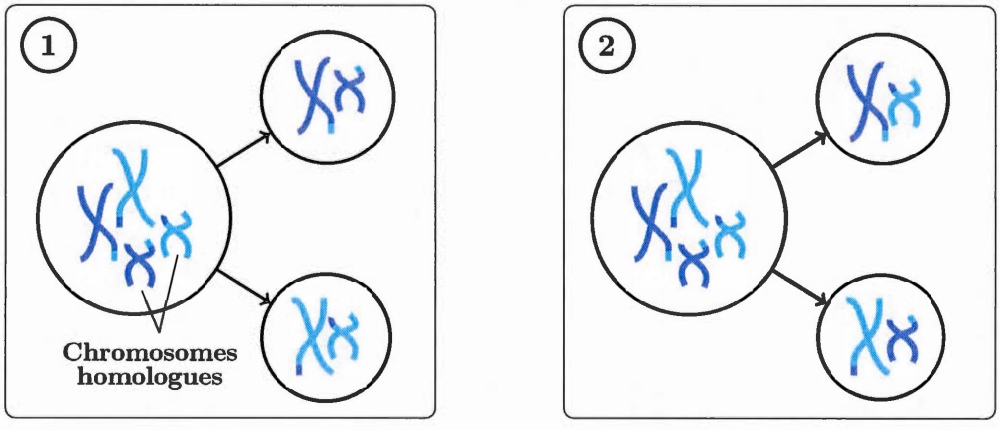


Figure 1.5: Exemple de deux types de séparations possibles pendant la première division de la méiose pour une cellule mère de deux paires de chromosomes homologues. Il y a donc $2^2 = 4$ sortes de cellules filles possibles. Ici, chaque chromosome est constitué de deux chromatides identiques.

Nous remarquons que pendant la méiose, un chromosome risque d'être séparé en deux morceaux, et que ces derniers se rattachent aux morceaux de son homologue (voir Figure 1.4). De ce fait, il est possible que deux gènes d'un même chromosome proviennent de deux chromosomes différents. Il s'agit d'une *recombinaison* génétique. Par conséquent, on observe un événement de recombinaison entre deux gènes lorsqu'un nombre impair d'enjambements survient entre ces deux gènes.

D'après les études de Thomas Hunt Morgan au début du 20^{e} siècle, la probabilité d'avoir un tel événement est proportionnelle à la distance entre leur locus. Ainsi, une recombinaison a plus de chance de survenir entre deux loci plus éloignés ; si cette chance est équivalente à 1%, on dit que la *distance génétique* entre ces deux loci est de 1 *centimorgan* (cM). La distance génétique entre deux gènes pourrait être estimée pour des petites distances par leur distance physique selon l'équation suivante :

$$1 \text{ cM} \approx 1 \text{ Mb}, \tag{1.1}$$

où 1 Mb correspond à 1 million de paires de bases.

Une autre source importante de la variabilité génétique est la *mutation* génétique, qui est une modification de l'ADN. Cette modification peut se produire lors de la mitose ou de la méiose. Une mutation peut donc apparaître dans les gamètes ou les cellules somatiques. Elle peut survenir pendant la réplication de l'ADN en touchant une ou plusieurs paires de bases azotées d'un gène, et il existe 3 formes possibles : substitution, insertion et délétion. La première forme correspond au remplacement d'une paire de bases par l'autre ; la deuxième et la troisième forme indiquent l'addition et la perte d'une ou plusieurs paires de bases. Il est possible qu'une mutation n'ait pas d'effet sur l'individu. Si cette mutation se trouve dans les cellules somatiques d'un individu, ses descendants ne porteront pas cette mutation. Par contre, si la mutation a lieu dans les gamètes, tous ses descendants vont la porter et certains pourront être affectés par celle-ci. Dans ce cas, il s'agit d'un caractère génétique. Dans la suite on se réfère à des individus atteints (cas) et non atteints (témoins) du caractère génétique.

1.3 Marqueurs génétiques

Deux personnes non-reliées ont en moyenne 99.9% de séquences d'ADN identiques et 0.1% de différences. Ces différences représentent donc les variations génétiques entre des individus. Un marqueur génétique est un gène ou une séquence polymorphe d'ADN qui montre ces variations, et sa localisation est parfaitement connue. La *cartographie génétique* est une méthode qui permet de déterminer l'emplacement des gènes influençant un caractère d'intérêt comme une maladie génétique, à l'aide de marqueurs génétiques. Lorsqu'il s'agit d'une recherche de l'emplacement sur une courte séquence, nous parlons de cartographie génétiques fine. Dans ce mémoire, on n'utilise qu'un seul type de marqueur génétique qui s'appelle *polymorphisme nucléotide simple* (de l'anglais *single-nucleotide polymorphism*, SNP en version abrégée). Parmi toutes les variations génétiques entre les individus, plus de 90% sont sous la forme de SNPs. Un SNP décrit une variation due à une seule paire de bases du génome. Par exemple, à une certaine position de la séquence, une partie des personnes dans une population a l'allèle A tandis que le reste a l'allèle C. Ces deux possibilités sur un même locus forment les deux allèles d'un SNP (voir Figure 1.6 - SNP1 ou SNP2).

La fréquence d'un SNP est mesurée par la fréquence de son allèle le plus rare ou mineur, c'est-à-dire, *minor allele frequency* en anglais (MAF en version abrégée). Les SNPs les plus courants sont ceux qui ont au moins 5% de MAF dans une population. Ce type de SNPs représente environ 90% de l'ensemble de SNPs. Parmi les 10% de SNPs les moins courants, ceux avec une MAF de 0.5% ou moins sont dits rares ; ceux avec une MAF de 0.5% à 5% sont dits peu courants. Un projet international spécifié qui étudie ces trois types de SNPs sera présenté dans le chapitre 4. Un SNP peut posséder deux ou plusieurs allèles. Mais en réalité, la plupart des SNPs ont seulement deux allèles comme illustré à la figure 1.6. En effet,

le taux de mutation sur une base azotée spécifique du génome est extrêmement faible (la taille du génome humain est d'environ 3 milliards paires de bases) et la probabilité d'avoir deux mutations avec une fréquence de 1% à la même position est presque négligeable.

Dans la suite du mémoire, on étudiera le lien entre les SNPs et un caractère génétique avec deux méthodes différentes. L'échantillon qu'on utilisera sera constitué de séquences de SNPs provenant d'individus non-reliés.

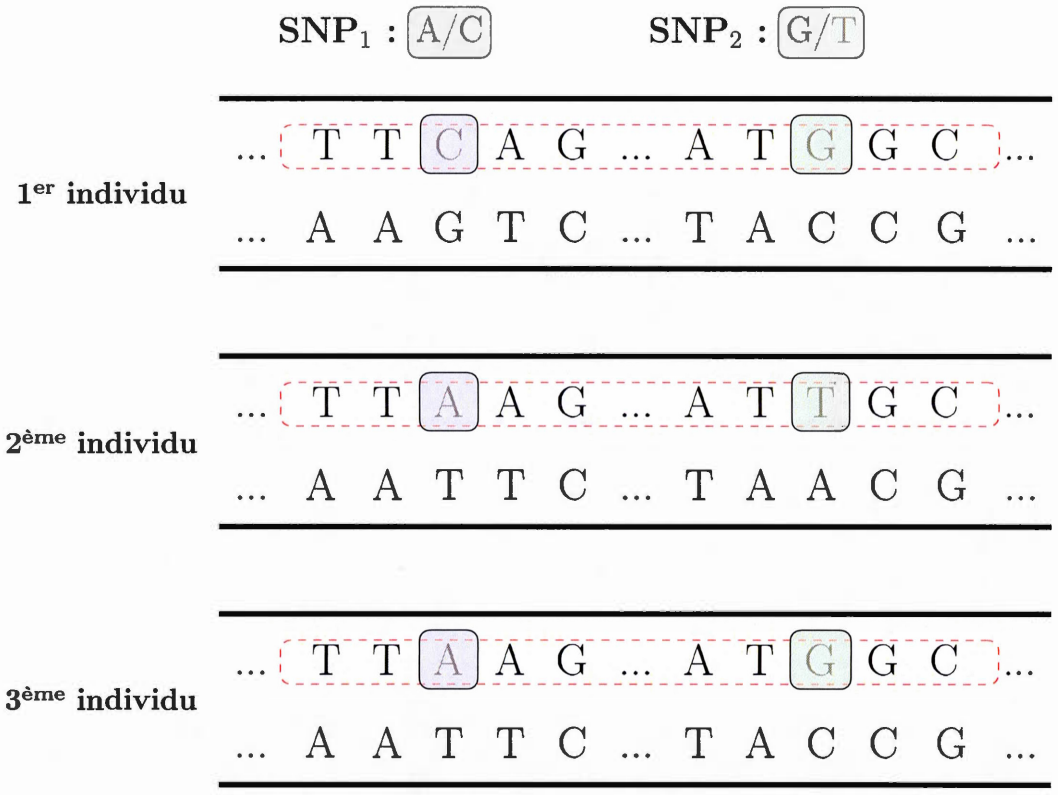


Figure 1.6: Exemple simple de deux SNPs. Il s'agit de deux variations sur une courte séquence d'ADN provenant de la même région chromosomique (endroits encadrés en pointillés) chez 3 personnes non reliées. Les séquences d'ADN sont identiques dans ces chromosomes, sauf deux bases surlignées qui montrent des variations. La première variation à gauche consiste en deux allèles A et C, et la deuxième variation à droite consiste en deux allèles G et T.

CHAPITRE II

THÉORIE DE LA COALESCENCE

La première méthode qu'on utilise dans ce mémoire pour étudier le lien entre les SNPs et un caractère génétique est basée sur la théorie de la coalescence. Cette dernière est une approche rétrospective qui simule la généalogie d'un échantillon d'une grande population. Nous allons décrire cette théorie en partant du modèle de Wright-Fisher ; ensuite, on présente la construction du graphe de recombinaison ancestral qui est une généralisation de la théorie de coalescence afin d'inclure la recombinaison génétique.

2.1 Modèle de Wright-Fisher

2.1.1 Présentation

Le modèle de Wright-Fisher a été développé par Fisher (1923) et Wright (1931). Il modélise la transmission des gènes d'une génération à une autre. C'est le modèle stochastique le plus utilisé pour étudier l'évolution génétique d'une population. Ce modèle fait les suppositions suivantes :

— La population a une taille finie et constante : $2N$ individus haploïdes, donc chaque individu a un seul parent ;

— les générations sont disjointes : cela veut dire que tous les individus d'une

même génération naissent (de la génération précédente) et décèdent au même moment ;

— il n'y a pas de sélection naturelle : tous les individus d'une même génération ont la même chance d'avoir des enfants.

Dans notre contexte chaque individu est représenté par une séquence génétique, c'est-à-dire une suite de SNPs.

2.1.2 Simulation de l'évolution

Sous le modèle Wright-Fisher, chaque séquence génétique dans une génération a une probabilité de $1/(2N)$ d'être l'enfant d'une séquence génétique spécifique de la génération précédente. Ainsi, pour générer une population qui suit le modèle de Wright-Fisher, si la taille de la génération est de $2N$ séquences génétiques au temps t, on simule une nouvelle génération de même taille au temps $t+1$ en effectuant $2N$ tirages avec remise des $2N$ séquences génétiques au temps t. Ainsi, le nombre d'enfants d'un parent provenant d'une génération de $2N$ séquences génétiques est une variable aléatoire qui suit une loi $Bin\,(2N, 1/(2N))$.

La figure 2.1 nous montre l'évolution d'une population de $2N = 6$ séquences génétiques pendant 6 générations selon le modèle Wright-Fisher. Chaque cercle représente une séquence génétique, où x_{ij} $(i = 0, \ldots, 5$ et $j = 1, \ldots, 6)$ représente la $j^{\text{ème}}$ séquence génétique à la $i^{\text{ème}}$ génération. Les séquences génétiques reliées par des arêtes ont des liens de parenté. Chaque séquence a un parent et des descendants sauf ceux de la première génération qui ont seulement des descendants et ceux de la dernière génération qui ont seulement un parent. Prenons la séquence génétique x_{04} et ses descendants comme exemple ; alors, x_{04} donne naissance à 1 enfant (x_{12}) au temps 1, et celui-ci donne naissance à 2 enfants (x_{23} et x_{25}) au temps 2, et ainsi de suite.

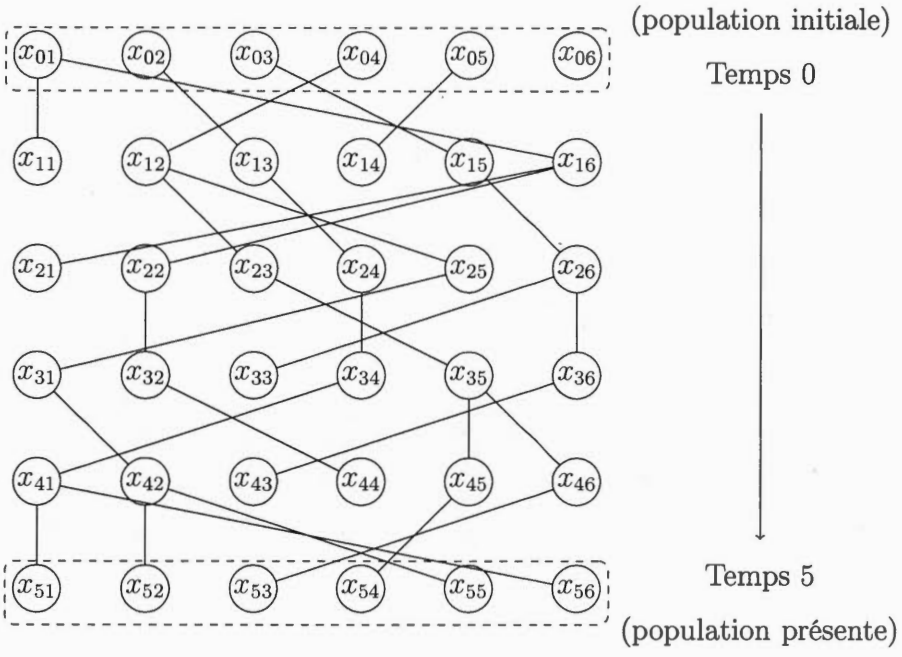


Figure 2.1: Exemple de l'évolution d'une population de $2N = 6$ séquences génétiques sous le modèle Wright-Fisher. La figure représente une simulation de cette population pendant 6 générations.

2.1.3 Étude de la généalogie

Une fois que l'évolution d'une population est créée, on peut trouver la généalogie des séquences génétiques de la population présente, c'est-à-dire les parents, grands-parents. On étudie la généalogie de la population en remontant dans le temps. Si l'on regarde dans l'ordre de construction de l'évolution (du passé vers le présent), on remarque que le nombre d'enfants d'une séquence influence le nombre d'enfants d'une autre séquence de la même génération. Par contre, si l'on regarde du présent vers le passé, le nombre de parents pour chaque séquence génétique est toujours 1, car on est dans un contexte d'une population haploïde. Donc le choix de parent d'une séquence n'influence pas le choix de l'autre d'une même génération. De plus,

en réalité la plupart des données qu'on peut obtenir viennent d'une population du moment présent.

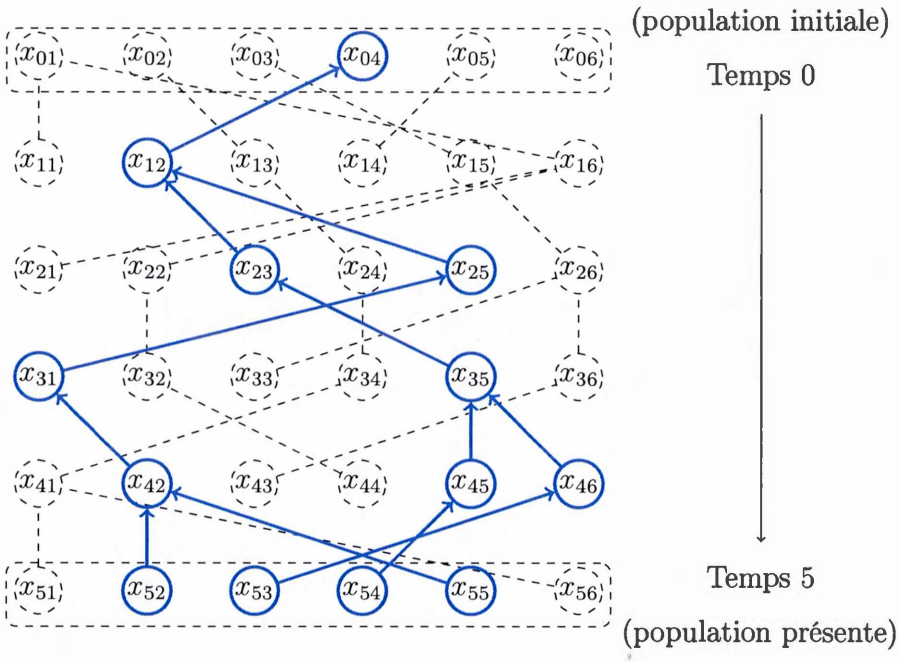


Figure 2.2: Exemple d'une généalogie de 4 séquences génétiques de la population présente. Cette généalogie est formée par les cercles encerclés en bleu et les arêtes qui les relient.

En suivant l'exemple précédent, la figure 2.2 nous montre un exemple de généalogie de 4 séquences génétiques de la population présente (en bleu). Dans cette généalogie, il y a deux séquences génétiques qui sont des ancêtres communs des 4 séquences considérées au temps 5 : x_{04} et x_{12}. Mais dans une étude de coalescence, on considère seulement la séquence génétique x_{12} au temps 1 qu'on appelle l'*ancêtre commun le plus récent*, le MRCA, de l'anglais «*most recent common ancestor*». L'évolution d'une séquence génétique, par exemple, x_{52} jusqu'à la séquence génétique x_{04} s'appelle lignée. La lignée d'une séquence génétique est donc un chemin qui nous montre tous ses ancêtres à des générations successives.

Clairement, une fois que deux lignées se rejoignent, elles restent confondues pour toutes les générations précédentes. Les informations nécessaires pour étudier une généalogie sont le nombre de générations pour atteindre un MRCA et les séquences génétiques qui ont trouvé un ancêtre commun. La généalogie peut donc se représenter comme dans la figure 2.3.

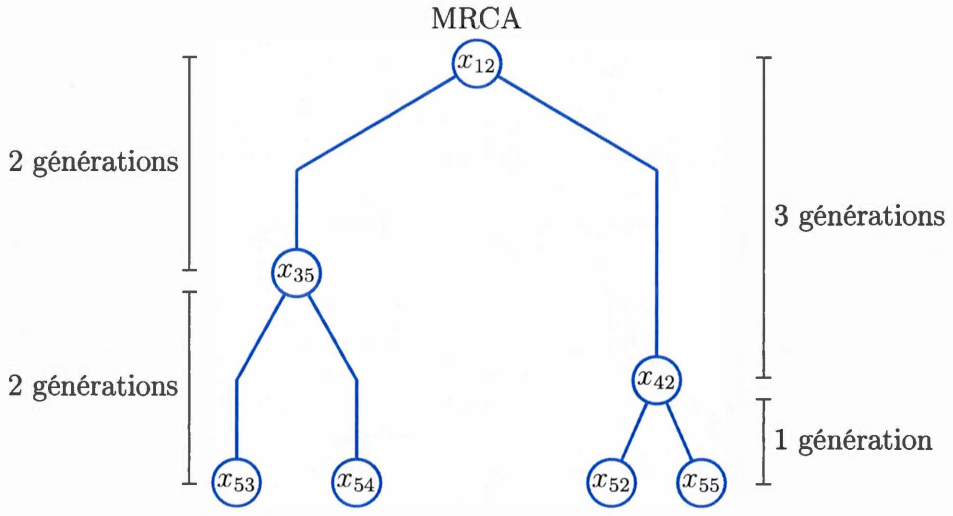


Figure 2.3: Illustration d'une généalogie qui est équivalente à celle dans la figure 2.2.

2.2 Théorie de la coalescence

Avant les années 1980, des scientifiques avaient déjà développé des approches qui permettaient de prédire l'évolution d'une population spécifique, mais ces approches ont des restrictions. Premièrement, elles sont toutes prospectives et cela ne permet pas de bien étudier l'histoire généalogique d'une population donnée ; deuxièmement, elles étudient sur tous les individus d'une population, tandis qu'en réalité on ne peut qu'obtenir des informations d'une petite partie de la population (Kingman, 2000). Au début des années 1980, une nouvelle approche est apparue :

la théorie de la coalescence (Kingman, 1982a,b).

L'événement de coalescence correspond au moment où deux lignées trouvent un ancêtre commun. Par exemple, on dit que les séquences génétiques x_{53} et x_{54} dans la figure 2.3 coalescent après deux générations, et x_{35} est la séquence génétique résultante. La théorie mathématique de la coalescence a été présentée par Kingman (1982a), et on l'appelle aussi le coalescent de Kingman. Cette théorie est un processus stochastique qui se base sur le modèle de Wright-Fisher quand la taille de population est grande. Elle vise à reconstruire l'histoire généalogique des séquences génétiques observées dans un échantillon de taille n provenant d'une population de taille $2N$ où $n \ll 2N$, jusqu'à ce qu'on trouve le MRCA de ces séquences génétiques. Il s'agit d'un processus aléatoire dont les états sont des ensembles de séquences génétiques à un temps donné. Lorsque deux séquences génétiques coalescent, on passe d'un état à un autre. De ce fait, on s'intéresse aussi à calculer le temps nécessaire pour avoir un événement de coalescence quand on veut générer un processus de coalescence.

2.2.1 Processus en temps discret

• Coalescence de 2 lignées

Cherchons la probabilité d'avoir le même parent pour deux séquences génétiques. Dans le modèle de Wright-Fisher, on a déjà vu que pour chaque séquence génétique d'une même génération, la probabilité d'être un enfant d'une séquence génétique donnée de la génération précédente est de $1/(2N)$. Ainsi, pour la deuxième séquence génétique, la probabilité de ne pas choisir le même parent que la première séquence est de $1 - 1/(2N)$. Par conséquent, la probabilité d'avoir deux grands-parents différents pour deux séquences génétiques est de $(1 - 1/(2N))^2$ (voir la figure 2.4). On peut donc déduire qu'après $k - 1$ générations, la probabilité de

ne pas avoir un ancêtre commun pour deux lignées est de $(1 - 1/(2N))^{k-1}$. Ainsi, après k générations, la probabilité d'avoir un ancêtre commun pour deux lignées est de $(1/(2N))(1 - 1/(2N))^{k-1}$. Donc le nombre de générations pour que deux lignées atteignent un ancêtre commun suit une loi géométrique de paramètre $1/(2N)$.

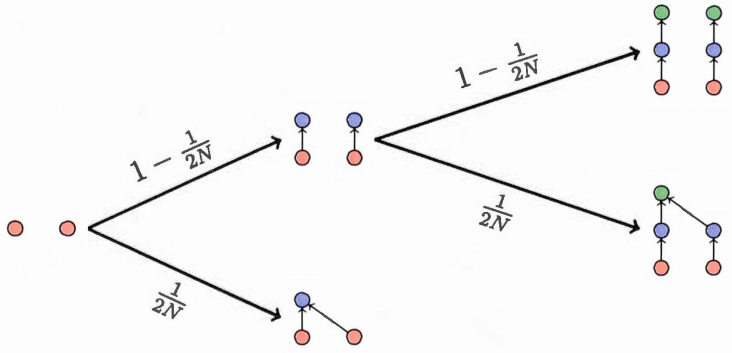


Figure 2.4: Calcul de la probabilité d'avoir des grands-parents différents pour deux séquences génétiques. Les cercles rouges, bleus et verts représentent respectivement les enfants, parents et grands-parents.

• Coalescence de n lignées

De la même manière, on peut calculer la probabilité de n'avoir aucune coalescence pour n lignées après k générations, c'est-à-dire la probabilité d'avoir n lignées distinctes après k générations, que l'on notera $P_n(k)$, $n = 2, \ldots, 2N$ et $k = 1, 2, \ldots$. Tout d'abord, cherchons la probabilité de n'avoir aucune coalescence pour n séquences génétiques après 1 génération. Dans ce cas, la première séquence génétique choisit 1 parent parmi $2N$, puis la deuxième choisit 1 parent parmi les $2N - 1$ restants, et la troisième choisit 1 parent parmi les $2N - 2$ restants, et

ainsi de suite. Cette probabilité s'écrit donc comme ci-dessous :

$$\begin{aligned}
P_n(1) &= \left(\frac{2N}{2N}\right)\left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)\ldots\left(\frac{2N-n+1}{2N}\right) \\
&= \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\ldots\left(1-\frac{n-1}{2N}\right) \\
&= \left(1-\frac{1}{2N}-\frac{2}{2N}+\frac{2}{(2N)^2}\right)\left(1-\frac{3}{2N}\right)\ldots\left(1-\frac{n-1}{2N}\right) \\
&= \left(1-\frac{1}{2N}-\frac{2}{2N}\right)\left(1-\frac{3}{2N}\right)\ldots\left(1-\frac{n-1}{2N}\right)+O\left(\frac{1}{N^2}\right) \\
&\vdots \\
&= 1-\frac{1}{2N}-\frac{2}{2N}-\frac{3}{2N}+\ldots-\frac{n-1}{2N}+O\left(\frac{1}{N^2}\right) \\
&= 1-\frac{1}{2N}\left(1+2+3+\ldots+(n-1)\right)+O\left(\frac{1}{N^2}\right) \\
&= 1-\frac{1}{2N}\left(\frac{n(n-1)}{2}\right)+O\left(\frac{1}{N^2}\right) \\
&= 1-\frac{1}{2N}\binom{n}{2}+O\left(\frac{1}{N^2}\right), \qquad (2.1)
\end{aligned}$$

où $O(1/N^2)$ représente tous les termes qui sont divisés par N à la puissance 2 ou supérieure : $1/N^2$, $1/N^3$, etc. Ce terme représente aussi la probabilité d'avoir plus de deux séquences génétiques qui coalescent à la même génération. En supposant que $n \ll 2N$, ce terme est négligeable. Ainsi $P_n(1) \approx 1 - \binom{n}{2}/(2N)$. Et le complément de cette probabilité, $\binom{n}{2}/(2N)$, s'approche de la probabilité d'avoir un événement de coalescence entre 2 séquences génétiques parmi n après une génération.

Posons T_n le nombre de générations qu'il faut pour que deux lignées parmi n coalescent ; alors, la probabilité que $T_n = k$ est :

$$P(T_n = k) \approx \left(1-\binom{n}{2}\frac{1}{2N}\right)^{k-1}\binom{n}{2}\frac{1}{2N}. \qquad (2.2)$$

Évidemment, T_n suit approximativement une loi géométrique de paramètre

$\binom{n}{2}/(2N)$. Il faut donc en moyenne $2N/\binom{n}{2}$ générations pour que deux lignées parmi n coalescent, où $n = 2, \ldots, 2N$.

2.2.2 Approximation du processus en temps continu

Le processus défini à la section 2.2.1 s'approxime par un processus en temps continu ($N \to \infty$). Sachant que la variable discrète T_n suit approximativement une loi géométrique, on peut obtenir que :

$$P(T_n \geq t_n) \approx \left(1 - \binom{n}{2}\frac{1}{2N}\right)^{t_n}, \tag{2.3}$$

où t_n représente un nombre de générations, si l'on travaille en temps discret.

Maintenant définissons T_n^c une variable continue qui représente le temps d'attente continu pour que deux lignées parmi n coalescent une première fois. Exprimons t en fonction d'une unité de mesure de $2N$ générations, qui est le temps moyen pour que deux lignées trouvent un ancêtre commun. Ainsi, on pose $t_n = 2Nt$ et $T_n = 2NT_n^c$. Donc, $P(T_n \geq t_n)$ devient :

$$\begin{aligned} P(2NT_n^c \geq 2Nt) &= P(T_n^c \geq t) \\ &\approx \left(1 - \binom{n}{2}\frac{1}{2N}\right)^{2Nt} \\ &\approx e^{-\binom{n}{2}t}, \end{aligned} \tag{2.4}$$

en utilisant la propriété : $(1 - x/N)^{Nt} \approx e^{-xt}$ quand x est fixe et $N \to \infty$.

T_n^c suit donc approximativement une loi exponentielle de paramètre $\binom{n}{2}$. Ainsi, pour simuler un processus de coalescence d'un échantillon de n séquences génétiques, on génère le temps d'attente T_s^c jusqu'au prochain événement de coalescence parmi s lignées, avec $T_s^c \sim \exp\left(\binom{s}{2}\right)$ où $s = n, n-1, ..., 2$. Par la suite, on choisit aléatoirement deux lignées parmi s et on les fait coalescer. On répète ces deux étapes jusqu'à ce qu'on trouve une seule lignée. Une généalogie créée par

ce processus est aussi appelée arbre de coalescence. De ce fait, la généalogie dans la figure 2.3 peut aussi représenter un arbre de coalescence d'un échantillon de 4 séquences génétiques.

2.3 Ajout d'événements de diversité génétique dans le processus de coalescence

Le processus de coalescence de base ne montre pas la diversité génétique de la population, on a donc besoin d'ajouter cette dernière dans le processus pour s'adapter à la réalité. Comme mentionné dans le chapitre précédent, il existe deux types d'événements principaux qui expliquent la diversité génétique chez l'être humain : mutation et recombinaison.

2.3.1 Coalescence avec mutation

Commençons par ajouter les événements de mutation dans le processus de coalescence. Nous supposons que les mutations sont neutres car les probabilités de mort et de reproduction ne changent pas dans la population. Cela nous indique que le processus de mutation est indépendant du processus de coalescence, il suffit donc de superposer les mutations sur l'arbre de coalescence une fois que ce dernier est construit. Une façon de choisir des emplacements de mutation est d'utiliser le modèle des sites infinis (Kimura, 1969). Ce modèle suppose que chaque séquence génétique est constituée d'une suite infinie des sites, et l'emplacement de la mutation est choisi aléatoirement sur un des sites qui n'ont jamais eu une mutation. Le nombre de sites mutés est donc équivalent au nombre de mutations qui se sont produites sur l'arbre de coalescence. C'est le modèle le plus simple et qui approxime bien la réalité quand la séquence génétique est longue et le taux de mutation est bas.

Posons μ la probabilité d'avoir une mutation par séquence par génération

et t le temps (en unités de $2N$ générations). Dans la suite, on considère le processus stochastique en temps continu. Tout comme la méthode pour trouver la distribution du temps d'attente d'un événement de coalescence, on obtient que, le temps d'attente T_1^m avant qu'une lignée subisse une mutation suit la loi suivante :

$$\begin{aligned} P(T_1^m \geq t) &= (1-\mu)^{2Nt} \\ &\approx \left(1 - \frac{\theta}{4N}\right)^{2Nt} \\ &\approx e^{-\frac{\theta}{2}t}, \end{aligned} \tag{2.5}$$

en utilisant la propriété : $(1 - x/N)^{Nt} \approx e^{-xt}$ quand x est fixe et $N \to \infty$. Le paramètre $\theta = 4N\mu$ est en fait le taux de mutation de la population. T_1^m suit donc approximativement une loi exponentielle de paramètre $\theta/2$. D'après le lien entre le processus de Poisson et la loi exponentielle, on considère que le nombre de mutations qui se produisent sur une branche de longueur l d'un arbre de coalescence suit une loi de Poisson d'espérance $\theta l/2$. De ce fait, l'ajout des mutations sur un arbre de coalescence se réalise par un processus de Poisson le long des branches, et l'espérance du nombre de mutations par branche est proportionnelle à la longueur de cette branche. Cette propriété servira dans le prochain chapitre.

Combinons les événements de coalescence avec les événements de mutation pour n lignées. Posons T_n^m le temps d'attente d'un événement de mutation sur une lignée parmi n, cette variable suit donc une loi exponentielle de paramètre $n\theta/2$, ce qui représente en fait le taux de mutation de n lignées. Quant à T_n^c, le temps d'attente d'un événement de coalescence entre deux lignées parmi n, il suit une loi exponentielle de paramètre $\binom{n}{2}$ comme démontré à la section précédente. En supposant que les événements de coalescence et de mutation sont indépendants, le temps d'attente avant qu'un de ces deux événements arrive suit une loi

exponentielle de paramètre :

$$\binom{n}{2} + \frac{n\theta}{2} = \frac{n(n-1+\theta)}{2}. \tag{2.6}$$

La probabilité d'avoir un événement de coalescence en premier est :

$$\frac{\binom{n}{2}}{n(n-1+\theta)/2} = \frac{n-1}{n-1+\theta}, \tag{2.7}$$

et la probabilité d'avoir un événement de mutation en premier est :

$$\frac{n\theta/2}{n(n-1+\theta)/2} = \frac{\theta}{n-1+\theta}. \tag{2.8}$$

Les résultats des équations (2.6), (2.7) et (2.8) sont des conséquences des propriétés classiques suivantes.

Proposition. *Soient X et Y deux variables aléatoires indépendantes suivant respectivement des lois exponentielles de paramètres λ_1 et λ_2.*

Propriété I :

La variable $min(X, Y)$ suit aussi une loi exponentielle, de paramètre $\lambda_1 + \lambda_2$.

Propriété II :

$$P(X < Y) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

En résumé, pour décrire un processus de coalescence avec des événements de mutation d'un échantillon de n séquences génétiques, on peut faire appel à une simulation avec les étapes suivantes en posant $s = n$:

Étape 1 : Simuler le temps d'attente avant le prochain événement comme une loi exponentielle de paramètre $s(s-1+\theta)/2$.

Étape 2 : Choisir le type du prochain événement :

— une coalescence, avec probabilité $(s-1)/(s-1+\theta)$;

— une mutation, avec probabilité $\theta/(s-1+\theta)$.

Étape 3 : Simuler le prochain événement :

— si l'événement choisi est une coalescence, sélectionner aléatoirement deux lignées parmi s et les faire coalescer, mettre $s = s-1$;

— si l'événement choisi est une mutation, sélectionner d'abord aléatoirement une lignée parmi s afin d'y insérer la mutation ; sélectionner ensuite aléatoirement un site non muté sur la séquence descendante de cette lignée et le faire muter.

Étape 4 : Si $s > 1$, retourner à l'étape 1.

Dénotons par «1» l'allèle muté et par «0» l'allèle non-muté sur chaque site. La figure 2.5 nous montre un processus de coalescence avec mutations pour un échantillon de 4 séquences génétiques. Dans ce processus, on a eu deux événements de mutation. D'après le modèle des sites infinis, deux sites devront produire ces deux mutations. On remarque aussi que l'échantillon obtenu suite à l'ajout de mutations change. Dans la section suivante, on présente un algorithme qui permet de générer une généalogie qui est consistante avec notre échantillon.

2.3.2 Coalescence avec recombinaison

L'ajout des événements de recombinaison dans le processus de coalescence avec mutation est réalisé en adaptant le premier modèle de coalescence avec recombinaison (Hudson, 1983). Rappelons que la population sous le modèle de Wright-Fisher est constituée de $2N$ haploïdes, tandis que les recombinaisons peuvent arriver seulement chez des cellules germinales, c'est-à-dire des cellules diploïdes. Mais cela n'est pas problématique car cette population peut être considérée comme une population de N diploïdes.

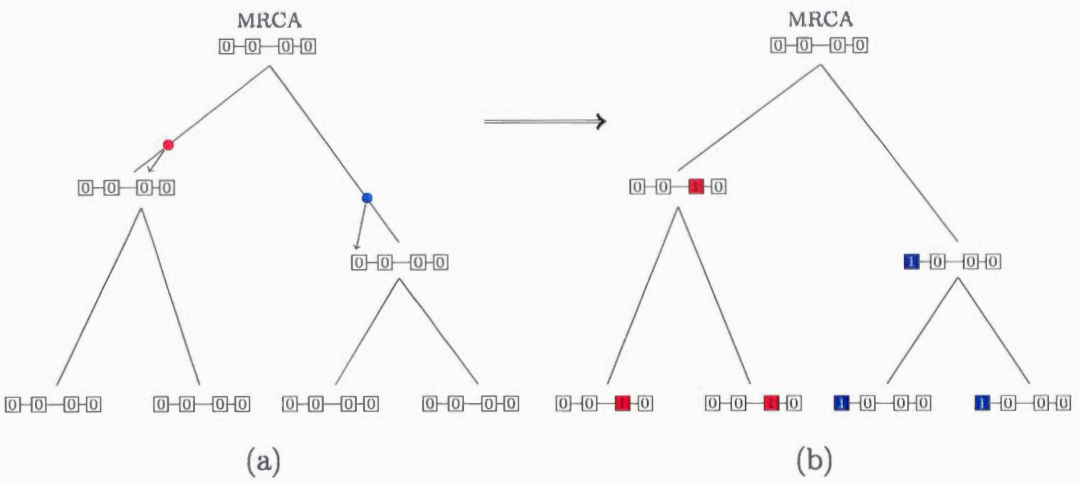


Figure 2.5: Illustration d'un processus de coalescence avec mutation pour un échantillon de 4 séquences génétiques. La figure (a) indique une simulation du processus de coalescence avec l'ajout de deux mutations sur deux lignées différentes ; les deux cercles colorés sont les deux mutations à apposer, et la position d'allèle à muter pour chaque mutation est choisie aléatoirement parmi les sites qui n'ont jamais eu une mutation. La figure (b) représente le changement d'allèle sur les séquences génétiques suite à l'ajout des mutations ; chaque case colorée représente un site où se situe l'allèle qui a hérité de la mutation.

Basé sur le modèle de Hudson, pour simuler un événement de recombinaison d'une séquence génétique rétrospectivement, on choisit d'abord un point de recombinaison sur cette séquence, et ensuite on crée deux séquences génétiques parentales dont l'une contient le matériel génétique à la gauche du point de recombinaison et l'autre contient le matériel génétique à la droite du point de recombinaison. La figure 2.6 illustre ce phénomène. Les cases blanches avec «0» représentent les marqueurs qui possèdent des allèles primitifs (non mutés), tandis que les cases noires avec «1» représentent les marqueurs qui possèdent des allèles dérivés (mutés). Ces deux types de marqueurs sont nommés marqueurs ancestraux,

dont les allèles constituent les séquences génétiques de notre échantillon et forment le matériel appelé matériel ancestral. Les cases grises avec « ?» représentent les marqueurs qui ont du matériel génétique inconnu. Ce dernier est appelé matériel non-ancestral, et le marqueur associé est nommé marqueur non-ancestral. Comme notre échantillon ne contient pas ce type de matériel, son existence dans la généalogie n'est pas informatif et on peut l'ignorer.

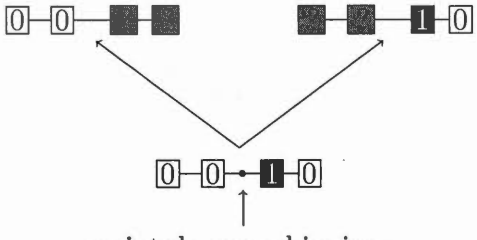


Figure 2.6: Représentation d'un événement de recombinaison sous le modèle de Hudson. Cet événement est simulé d'une manière rétrospective, les allèles inconnus des séquences génétiques parentales sont présentés par des points d'interrogation.

Afin de procéder à l'ajout de recombinaison dans un processus de coalescence avec mutation, on définit par r la probabilité d'avoir une recombinaison par séquence par génération et t le temps exprimé en unités de $2N$ générations. Semblable à la méthode pour trouver la distribution du temps d'attente continu d'un événement de coalescence ou de mutation, la distribution du temps d'attente continu T^r d'un événement de recombinaison sur une lignée s'obtient comme ci-dessous :

$$\begin{aligned} P(T^r \geq t) &= (1-r)^{2Nt} \\ &\approx \left(1 - \frac{\rho}{4N}\right)^{2Nt} \\ &\approx e^{-\frac{\rho}{2}t}, \end{aligned} \tag{2.9}$$

où le paramètre $\rho = 4Nr$ représente le taux de recombinaison de la population, et les deux dernières équations sont équivalentes grâce à la propriété : $(1 - x/N)^{Nt} \approx$

e^{-xt} quand x est fixe et $N \to \infty$. Ainsi, T^r suit une loi exponentielle de paramètre $\rho/2$.

À présent, associons les événements de recombinaison et le processus de coalescence avec mutation pour n lignées. De manière analogique du calcul de la sous section précédente, en sachant que les événements de recombinaison sont indépendants de ceux de coalescence et de mutation, le temps d'attente avant qu'un de ces trois événements arrive suit une loi exponentielle de paramètre :

$$\binom{n}{2} + \frac{n\theta}{2} + \frac{n\rho}{2} = \frac{n(n-1+\theta+\rho)}{2}. \tag{2.10}$$

La probabilité d'avoir un événement de coalescence en premier est :

$$\frac{\binom{n}{2}}{n(n-1+\theta+\rho)/2} = \frac{n-1}{n-1+\theta+\rho}, \tag{2.11}$$

la probabilité d'avoir un événement de mutation en premier est :

$$\frac{n\theta/2}{n(n-1+\theta+\rho)/2} = \frac{\theta}{n-1+\theta+\rho}, \tag{2.12}$$

et la probabilité d'avoir un événement de recombinaison en premier est :

$$\frac{n\rho/2}{n(n-1+\theta+\rho)/2} = \frac{\rho}{n-1+\theta+\rho}. \tag{2.13}$$

Afin de simuler un processus de coalescence avec des événements de mutation et de recombinaison pour un échantillon de n séquences génétiques, on pourrait imaginer l'algorithme suivant en posant $s = n$:

Étape 1 : Simuler le temps d'attente avant le prochain événement avec une loi exponentielle de paramètre $s(s-1+\theta+\rho)/2$.

Étape 2 : Choisir le type du prochain événement :

- une coalescence, avec probabilité $(s-1)/(s-1+\theta+\rho)$;
- une mutation, avec probabilité $\theta/(s-1+\theta+\rho)$;

— une recombinaison, avec probabilité $\rho/(s-1+\theta+\rho)$.

Étape 3 : Simuler le prochain événement :

— si l'événement choisi est une coalescence, sélectionner aléatoirement deux lignées parmi s et les faire coalescer, mettre $s = s-1$;

— si l'événement choisi est une mutation, sélectionner d'abord aléatoirement une séquence pouvant muter parmi s ; sélectionner ensuite aléatoirement un site pouvant muter sur cette séquence pour apposer la mutation : ce site passe alors du statut mutant au statut primitif ;

— si l'événement choisi est une recombinaison, sélectionner d'abord aléatoirement une séquence parmi s ; sélectionner ensuite aléatoirement un point de recombinaison sur celle-ci ; créer à la fin deux séquences parentales dont chacune contient respectivement une partie du matériel ancestral de la séquence à recombiner, et mettre $s = s+1$.

Étape 4 : Si $s > 1$, retourner à l'étape 1.

Ainsi, tel que décrit à l'étape 3 de cet algorithme, dans la reconstruction d'une généalogie en remontant dans le temps, les mutations qui se sont produites en avançant dans le temps sont « supprimées » en remontant dans le temps. De ce fait, si une mutation se produit, celle-ci se produit à un site sur une séquence particulière, telle que cette séquence est la seule de l'échantillon au temps où se produit cet événement possédant cette mutation : ce site change alors de statut, il passe de l'allèle muté à l'allèle non-muté.

La figure 2.7 illustre un exemple de ce processus pour un échantillon de 4 séquences génétiques. Une généalogie créée par un tel processus s'appelle

graphe de recombinaison ancestral, que l'on notera ARG de l'anglais «*ancestral recombination graph*».

Avec l'algorithme présenté on a les aspects suivants qu'on reprendra au prochain chapitre :

— grâce à la nouvelle méthode de superposition des mutations sur la généalogie, l'ARG obtenu est toujours consistant avec notre échantillon ;

— l'événement de coalescence peut se passer entre deux séquences génétiques différentes si leurs portions différentes sont non-ancestrales (voir les événements C_1 et C_2 sur la figure 2.7) ;

— le MRCA est toujours ancestral et primitif, il est donc supposé connu.

Un ARG peut s'exprimer aussi par une superposition des *arbres partiels* de tous les marqueurs ; un arbre partiel est la généalogie d'un seul marqueur dont l'allèle est ancestral. La figure 2.8 illustre un arbre partiel de l'ARG dans la figure 2.7 pour le premier marqueur.

2.4 Calcul des probabilités des événements en considérant la présence du matériel non-ancestral

Puisque le matériel non-ancestral est non-informatif, lors de la construction d'un ARG, on doit modifier la probabilité d'avoir un événement donné pour tenir compte de la présence de ce type de matériel. Parmi les trois types d'événements, les événements de mutation et de recombinaison sont influencés par la présence de matériel non-ancestral, car on doit choisir un emplacement sur la séquence génétique où se produira l'événement, dans lequel seulement le matériel ancestral doit être présent.

Supposons qu'on est à une génération donnée avec n séquences génétiques et n_i représente le nombre de séquences de type i (avec mêmes allèles aux mêmes

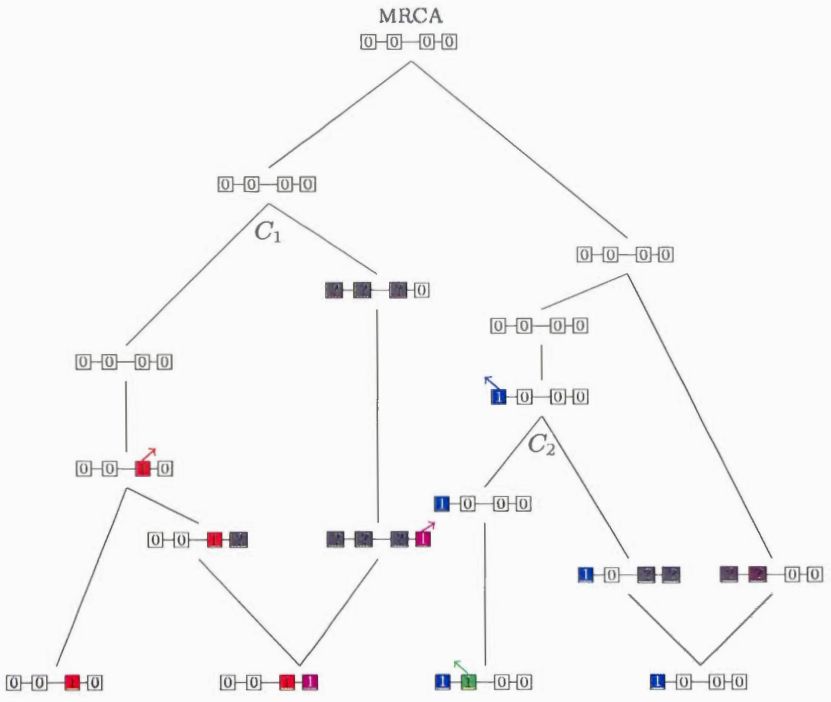


Figure 2.7: Illustration d'un graphe de recombinaison ancestral provenant d'un échantillon de 4 séquences génétiques. Cette figure représente un ARG où la superposition des mutations n'est pas représentée de la même façon que dans la figure 2.5. Les événements de coalescence peuvent se produire entre deux séquences génétiques différentes lorsque leurs parties différentes sont non-ancestrales (C_1 et C_2).

positions), donc $\sum_i n_i = n$. Posons a_i le nombre de marqueurs ancestraux d'une séquence génétique de type i et r_i la distance entre deux marqueurs qui se situent sur les deux bouts de la partie ancestrale d'une séquence génétique de type i. La proportion des marqueurs ancestraux peut donc être définie par l'équation suivante :

$$\alpha = \frac{\sum_i n_i a_i}{nL}, \tag{2.14}$$

où L est le nombre total de marqueurs d'une séquence génétique à L marqueurs.

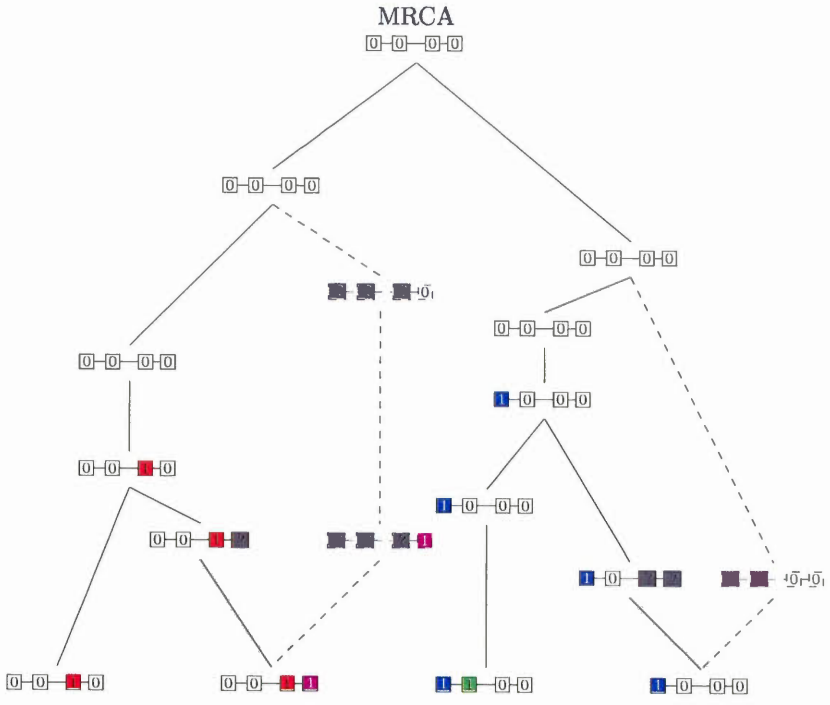


Figure 2.8: Exemple d'un arbre partiel de l'ARG dans la figure 2.7 pour le premier marqueur. Les branches et séquences génétiques qui ne sont pas en pointillé forment l'arbre partiel du premier marqueur.

Ainsi, le taux de mutation de ces n séquences génétiques devient $n\alpha\theta/2$. De même, la proportion des longueurs de séquences génétiques comprises entre les marqueurs ancestraux peut être définit comme ci-dessous :

$$\beta = \frac{\sum_i n_i r_i}{nr}, \tag{2.15}$$

où r est la longueur totale d'une séquence génétique. Le nouveau taux de recombinaison de ces n séquences génétiques est $n\beta\rho/2$. Par conséquent, en considérant la présence du matériel non-ancestral, les probabilités que le prochain événement soit une coalescence, une mutation ou une recombinaison deviennent

respectivement :

$$P(C) = \frac{n-1}{n-1+\alpha\theta-\beta\rho}, \tag{2.16}$$

$$P(M) = \frac{\alpha\theta}{n-1+\alpha\theta-\beta\rho}, \tag{2.17}$$

$$P(R) = \frac{\beta\rho}{n-1+\alpha\theta-\beta\rho}. \tag{2.18}$$

Cette remarque sera utilisée dans la méthode DMap du prochain chapitre, où la présence du matériel non-ancestral joue un rôle important.

CHAPITRE III

MÉTHODE DE CARTOGRAPHIE GÉNÉTIQUE FINE VIA LE PROCESSUS DE COALESCENCE : MÉTHODE DMAP

Comme toutes les méthodes de cartographie génétique, la méthode DMap (Descary, 2012) sert à estimer la position d'une mutation qui influence un caractère génétique. Cette méthode a été développée spécialement pour un caractère qui ne s'exprime pas totalement quand le génotype associé est présent. C'est une méthode qui combine des points forts de différentes méthodes de cartographie génétique basées sur les généalogies de population, comme MapARG (Larribe *et al.*, 2002) et Margarita (Minichiello et Durbin, 2006).

3.1 Contexte général

Présentons d'abord la forme des données utilisées par la méthode DMap, elle est décrite ci-dessous :

— l'échantillon est un ensemble des courtes séquences génétiques et chaque séquence est considérée comme une suite de SNP ;

— chaque séquence génétique provient d'un individu avec un phénotype connu d'un caractère génétique d'intérêt ;

— le caractère d'intérêt est relié uniquement à une mutation, que l'on appelle TIV (*trait influencing variant* en anglais), son allèle et sa position dans

l'échantillon sont inconnus.

Afin de décrire cette méthode, introduisons d'abord quelques termes et notations pour le contexte d'une population d'haploïdes. Posons «A» le type d'allèle de TIV qui est associé au caractère d'intérêt. A=0 correspond à l'allèle non-muté et A=1 correspond à l'allèle muté. La séquence génétique qui possède l'allèle «0» est appelée non-porteuse de la mutation, tandis que la séquence génétique qui possède l'allèle «1» est appelée porteuse de la mutation. La proportion de séquences génétiques dans la catégorie porteuse (ou non-porteuse) de la mutation qui proviennent d'individus atteints par le caractère d'intérêt est la pénétrance, notée par F. Pour une population d'haploïdes, $F = (f_0, f_1)$, où f_0 représente la probabilité d'être atteint par le caractère d'intérêt quand on a 0 copie d'allèle muté et f_1 représente la probabilité d'être atteint par le caractère d'intérêt quand on a 1 copie d'allèle muté. Quand $f_0 \neq 0$, on parle de phénocopie ; une séquence génétique provenant d'un individu atteint par le caractère d'intérêt peut être non-porteuse de la mutation. Dans ce cas, le génotype n'est pas le seul facteur qui détermine le caractère, l'environnement peut aussi jouer un rôle. Si $f_1 \neq 1$, on parle alors de pénétrance incomplète ; une séquence génétique provenant d'un individu non atteint par le caractère d'intérêt peut être porteuse de la mutation. La pénétrance $F = (f_0, f_1)$ et la fréquence du caractère d'intérêt dans la population sont supposées connues dans la méthode DMap.

Pour trouver la position du TIV, la méthode DMap choisit d'abord certaines positions candidates de l'emplacement du TIV. Par la suite, la vraisemblance de ces positions sélectionnées est calculée, et l'emplacement du TIV est estimé par la position qui maximise la vraisemblance. Les sections suivantes nous décrivent en détails comment cela fonctionne.

3.2 Calcul de la vraisemblance

Supposons qu'il y a n_0 séquences génétiques toutes de même longueur r mégabases dans notre échantillon, dont chacune est formée de L marqueurs de positions connues. La position inconnue du TIV est représentée par c_T. Définissons $Y = \{y_1, y_2, ..., y_Z\}$ un vecteur des Z positions candidates (connues) et x_{y_z} le marqueur qui se situe à la position y_z sur une séquence génétique, où $z \in \{1, 2, ..., Z\}$. La position inconnue c_T sera donc estimée à une position y_z. Notons $\mathbf{H_0}$ l'ensemble des séquences génétiques de l'échantillon et $\Phi = (\phi_1, \phi_2, ..., \phi_{n_0})$ le vecteur des phénotypes de toutes les séquences génétiques, où $\phi_i, i \in \{1, 2, .., n_0\}$ représente le phénotype de la $i^{\text{ème}}$ séquence génétique. On obtient la fonction de vraisemblance suivante de la position c_T du TIV (notre paramètre à estimer) :

$$\begin{aligned} L(c_T) &= P(\mathbf{H_0}, \Phi \mid c_T) \\ &= \int P(\mathbf{H_0}, \Phi \mid G, c_T) \cdot P(G \mid c_T)\, dG \\ &= \int P(\mathbf{H_0} \mid \Phi, G, c_T) \cdot P(\Phi \mid G, c_T) \cdot P(G \mid c_T)\, dG, \end{aligned} \tag{3.1}$$

où G est une généalogie. En sachant que la probabilité de générer une généalogie qui est consistante avec notre échantillon est très faible, une bonne manière pour calculer l'équation (3.1) est d'utiliser la méthode d'échantillonnage préférentiel (*importance sampling* en anglais) avec une distribution instrumentale $Q(\cdot)$. La fonction de vraisemblance se réécrit de la façon suivante :

$$\begin{aligned} L(c_T) &= \int \frac{P(\mathbf{H_0} \mid \Phi, G, c_T) \cdot P(\Phi \mid G, c_T) \cdot P(G \mid c_T)}{Q(G)} \cdot Q(G)\, dG \\ &= E_Q \left[\frac{P(\mathbf{H_0} \mid \Phi, G, c_T) \cdot P(\Phi \mid G, c_T) \cdot P(G \mid c_T)}{Q(G)} \right]. \end{aligned} \tag{3.2}$$

L'équation (3.2) est une espérance, et son estimateur sans biais peut être donné par une moyenne de :

$$\frac{P(\mathbf{H_0} \mid \Phi, \mathbf{G}, c_T) \cdot P(\Phi \mid \mathbf{G}, c_T) \cdot P(\mathbf{G} \mid c_T)}{Q(\mathbf{G})}, \tag{3.3}$$

où $\mathbf{G}$ est un ensemble des généalogies dont chacune est générée selon une loi instrumentale $Q(\cdot)$.

C'est-à-dire, l'espérance (3.2) est approximée par

$$L(c_T) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(\mathbf{H}_0 \mid \Phi, G^{(i)}, c_T) \cdot P(\Phi \mid G^{(i)}, c_T) \cdot P(G^{(i)} \mid c_T)}{Q(G^{(i)})}, \tag{3.4}$$

où les généalogies $G^{(i)}$, $i = 1, ..., M$ sont simulées selon la distribution $Q(G)$.

Une bonne distribution instrumentale nous permet d'augmenter l'efficacité de l'approximation précédente, et la méthode DMap utilise celle proposée par Fearnhead et Donnelly (2001) ; cette distribution sera présentée dans la section 3.4 . Supposons que l'on génère toujours une généalogie consistante avec $\mathbf{H}_0$, ce qui veut dire que l'échantillon obtenu à partir de la généalogie simulée $G^{(i)}$ est toujours le même que l'échantillon présent $\mathbf{H}_0$; ainsi $P(\mathbf{H}_0 \mid G^{(i)}) = 1$, pour tout $i = 1, \cdots, M$. Le principe de la méthode permettant d'obtenir ce genre de généalogie est présenté à la sous-sections 2.3.2.

Dans la méthode DMap, on suppose que le TIV n'est pas inclus dans l'ensemble des séquences génétiques $\mathbf{H}_0$, donc $\mathbf{H}_0$ est indépendant de la position du TIV. Comme les généalogies $G^{(i)}$ sont générées en partant de $\mathbf{H}_0$, elles ne dépendent pas de la position du TIV, c'est-à-dire, $P(G^{(i)} \mid c_T) = P(G^{(i)})$. Ainsi, on peut supposer que $\mathbf{H}_0$ est indépendant du phénotype du caractère d'intérêt, donc $P(\mathbf{H}_0 \mid \Phi, G^{(i)}, c_T) = P(\mathbf{H}_0 \mid G^{(i)}) = 1$, pour tout $i = 1, \cdots, M$. Ainsi, l'équation (3.4) devient :

$$L(c_T) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(\Phi \mid G^{(i)}, c_T) \cdot P(G^{(i)})}{Q(G^{(i)})}. \tag{3.5}$$

Pour trouver le résultat final de cette équation, on présente dans les prochaines sections le calcul de chaque terme.

3.3 Calcul de la probabilité d'une généalogie $P(G)$

On commence par trouver la probabilité d'une généalogie générée. Dénotons $\mathbf{H}_\tau$, $\tau \in \{0, 1, \ldots, \tau^*\}$ le vecteur des séquences génétiques à la génération τ de la généalogie G, où 0 est l'indice de la génération présente et τ^* de la génération de l'ancêtre commun MRCA. La généalogie G se présente donc de la manière rétrospective avec les états suivants : $\mathbf{H}_0, \mathbf{H}_1, \ldots, \mathbf{H}_{\tau^*}$. Au moment où un événement de coalescence, de mutation ou de recombinaison se produit, on passe d'un état à un autre dans cette généalogie. Comme mentionné dans le chapitre 2, l'état d'une génération d'une généalogie dépend seulement de l'état de la génération précédente. Sous l'hypothèse que le MRCA est connu, en utilisant un raisonnement par récurrence, on obtient que la probabilité d'une généalogie est :

$$\begin{aligned}
P(G) &= P(\mathbf{H}_0, \mathbf{H}_1, \ldots, \mathbf{H}_{\tau^*}) \\
&= P(\mathbf{H}_0 \mid \mathbf{H}_1, \ldots, \mathbf{H}_{\tau^*}) \cdot P(\mathbf{H}_1, \ldots, \mathbf{H}_{\tau^*}) \\
&= P(\mathbf{H}_0 \mid \mathbf{H}_1) \cdot P(\mathbf{H}_1, \ldots, \mathbf{H}_{\tau^*}) \\
&= P(\mathbf{H}_0 \mid \mathbf{H}_1) \cdot P(\mathbf{H}_1 \mid \mathbf{H}_2) \cdot P(\mathbf{H}_2, \ldots, \mathbf{H}_{\tau^*}) \\
&\vdots \\
&= P(\mathbf{H}_{\tau^*}) \cdot \prod_{\tau=0}^{\tau^*-1} P(\mathbf{H}_\tau \mid \mathbf{H}_{\tau+1}) \\
&= \prod_{\tau=0}^{\tau^*-1} P(\mathbf{H}_\tau \mid \mathbf{H}_{\tau+1}), \qquad (3.6)
\end{aligned}$$

où $P(\mathbf{H}_\tau \mid \mathbf{H}_{\tau+1})$ représente la probabilité conditionnelle de l'état $\mathbf{H}_\tau$ en sachant l'état $\mathbf{H}_{\tau+1}$. Rappelons que l'on étudie la généalogie du présent vers le passé, donc le prochain événement (ou le nouvel état) signifie en fait l'événement (ou l'état) à la génération précédente. La probabilité conditionnelle à chercher dépend du type d'événement de transition entre les deux états ; pour la calculer, il faut donc distinguer les trois types d'événement possibles : coalescence, mutation et

recombinaison.

Selon les résultats du chapitre précédent, les probabilités d'avoir une coalescence, une mutation et une recombinaison lors du prochain événement sont respectivement :

$$P(C) = \frac{n-1}{n-1+\alpha\theta-\beta\rho}, \tag{3.7}$$

$$P(M) = \frac{\alpha\theta}{n-1+\alpha\theta-\beta\rho}, \tag{3.8}$$

$$P(R) = \frac{\beta\rho}{n-1+\alpha\theta-\beta\rho}, \tag{3.9}$$

où on a :

n : le nombre de séquences génétiques de l'état $\mathbf{H}_\tau$;

θ : le taux de mutation de la population ;

ρ : le taux de recombinaison de la population ;

α : la proportion des marqueurs ancestraux de $\mathbf{H}_\tau$;

β : la proportion des longueurs de séquences génétiques comprises entre les marqueurs ancestraux de $\mathbf{H}_\tau$, c'est-à-dire la proportion des segments de séquences génétiques où un événement de recombinaison peut se produire à l'état $\mathbf{H}_\tau$.

La probabilité conditionnelle d'intérêt sera calculée à l'aide de ces probabilités et de ces paramètres.

$\mathbf{H}_\tau \to \mathbf{H}_{\tau+1}$: coalescence

Supposons que l'événement de transition entre les deux états $\mathbf{H}_\tau$ et $\mathbf{H}_{\tau+1}$ soit une coalescence. Il faut considérer deux situations possibles : une coalescence entre deux séquences génétiques identiques et une coalescence entre deux types de séquences génétiques qui ont du matériel non ancestral différent. Notons C_{ii}^i une coalescence entre deux séquences identiques de type i résultant en une séquence

de type i et C_{ij}^{k} une coalescence entre deux séquences différentes i et j résultant en une séquence de type k. Dans ces deux cas, le nombre de séquences génétiques au nouvel état diminue, c'est-à-dire $\mid \mathbf{H}_{\tau+1} \mid = n-1$, et cette diminution est issue de la disparition d'une séquence génétique de l'état $\mathbf{H}_{\tau}$.

Utilisons $\mathbf{H}_{\tau+1} = \mathbf{H}_{\tau} \oplus C_{ii}^{i}$ pour indiquer le nouvel état $\mathbf{H}_{\tau+1}$ qui est obtenu par une coalescence C_{ii}^{i} en partant de l'état $\mathbf{H}_{\tau}$. Dans ce cas, c'est le nombre de séquences génétiques de type i qui diminue. Donc la probabilité conditionnelle $P(\mathbf{H}_{\tau} \mid \mathbf{H}_{\tau+1})$ est calculée par :

$$P(\mathbf{H}_{\tau} \mid \mathbf{H}_{\tau+1} = \mathbf{H}_{\tau} \oplus C_{ii}^{i}) = P(C) \cdot \frac{n_i - 1}{n-1}, \tag{3.10}$$

où $P(C)$ a été définie à la formule (3.7). La fraction $(n_i - 1)/(n-1)$ représente la probabilité qu'une séquence génétique de type i soit produite à l'état $\mathbf{H}_{\tau+1}$ lors de cet événement de coalescence, où n_i est le nombre de séquences génétiques de type i à l'état $\mathbf{H}_{\tau}$.

Utilisons $\mathbf{H}_{\tau+1} = \mathbf{H}_{\tau} \oplus C_{ij}^{k}$ pour représenter le nouveau état $\mathbf{H}_{\tau+1}$ qui est obtenu par une coalescence C_{ij}^{k} en partant de l'état $\mathbf{H}_{\tau}$. Dans cette situation, la probabilité conditionnelle recherchée devient :

$$P(\mathbf{H}_{\tau} \mid \mathbf{H}_{\tau+1} = \mathbf{H}_{\tau} \oplus C_{ij}^{k}) = P(C) \cdot \frac{n_k + 1 - \mathbb{1}_{\{i=k\}} - \mathbb{1}_{\{j=k\}}}{n-1}, \tag{3.11}$$

où $\mathbb{1}_{\{\cdot\}}$ est une fonction indicatrice, et le terme $(n_k + 1 - \mathbb{1}_{\{i=k\}} - \mathbb{1}_{\{j=k\}})/(n-1)$ représente la probabilité que la séquence génétique produite à l'état $\mathbf{H}_{\tau+1}$ à cause de cet événement de coalescence soit de type k. Notons que k peut être différent ou identique à i ou/et j.

$\mathbf{H}_{\tau} \rightarrow \mathbf{H}_{\tau+1}$: mutation

Notons $M_i^j(m)$ une mutation au marqueur m en passant du type de séquence génétique i à l'état $\mathbf{H}_{\tau}$ au type de séquence génétique j à l'état $\mathbf{H}_{\tau+1}$. Utilisons

$\mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus M_i^j(m)$ pour représenter ce type d'événement de transition. Ainsi, on obtient la probabilité conditionnelle par l'équation suivante :

$$P\left(\mathbf{H}_\tau \mid \mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus M_i^j(m)\right) = P(M) \cdot \frac{1}{\alpha L} \cdot \frac{n_j + 1}{n}. \tag{3.12}$$

où $P(M)$ a été définie à la formule (3.7). La fraction $1/(\alpha L)$ correspond à la probabilité d'avoir une mutation au marqueur m d'une séquence génétique de type i de l'état $\mathbf{H}_\tau$, où le dénominateur αL représente le nombre total des marqueurs ancestraux à cet état. Le terme $(n_j + 1)/n$ est la probabilité que la séquence génétique de type j à l'état $\mathbf{H}_{\tau+1}$ soit le résultat de cette mutation, car on a n_j séquences de type j à l'état $\mathbf{H}_\tau$ et $n_j + 1$ à l'état $\mathbf{H}_{\tau+1}$.

$\mathbf{H}_\tau \rightarrow \mathbf{H}_{\tau+1}$: recombinaison

Notons $R_i^{jk}(p)$ une recombinaison dans l'intervalle p d'une séquence génétique de type i à l'état $\mathbf{H}_\tau$, j et k soient les deux types de séquences génétiques produites à l'état $\mathbf{H}_{\tau+1}$. Supposons maintenant que l'événement de transition soit celui-ci. Cette sorte de transition est notée par $\mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus R_i^{jk}(p)$. La probabilité conditionnelle d'intérêt est obtenue par :

$$P(\mathbf{H}_\tau \mid \mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus R_i^{jk}(p)) = P(R) \cdot \frac{r_p}{\beta r} \cdot \frac{(n_j + 1)(n_k + 1)}{n(n+1)}, \tag{3.13}$$

où $P(R)$ a été définie à la formule (3.7). La fraction $r_p/\beta r$ représente la probabilité que le point de recombinaison soit situé dans l'intervalle p d'une séquence génétique de type i de l'état $\mathbf{H}_\tau$, où r_p est la longueur de cette intervalle. Le terme $(n_j + 1)(n_k + 1)/\left(n(n+1)\right)$ représente la probabilité que les deux types de séquences génétiques j et k soient produites à l'état $\mathbf{H}_{\tau+1}$ après cette recombinaison, car le nombre de séquence de type j et k ainsi que le nombre total de séquence augmente de 1.

3.4 La distribution instrumentale $Q(G)$

Par défaut, la méthode DMap utilise une distribution instrumentale proposée par Fearnhead et Donnelly (2001) pour générer des généalogies consistantes avec nos données. Selon eux, la meilleure distribution pour construire une généalogie s'écrit comme la suivante :

$$Q(\mathbf{H}_{\tau+1} \mid \mathbf{H}_\tau) = P(\mathbf{H}_\tau \mid \mathbf{H}_{\tau+1}) \cdot \frac{\pi(\mathbf{H}_{\tau+1})}{\pi(\mathbf{H}_\tau)}. \tag{3.14}$$

Cette distribution est représentée par la probabilité de transition quand l'on génère une généalogie du présent vers le passé. $\pi(\mathbf{H}_\tau)$ ici signifie la probabilité d'avoir un échantillon de séquences génétiques tirées aléatoirement d'une population qui est identique à l'ensemble de séquences à l'état $\mathbf{H}_\tau$, si l'on ne tient compte que du matériel ancestral de $\mathbf{H}_\tau$. Cette probabilité pourrait être calculée à travers $\pi(\cdot \mid \mathbf{H}_\tau \backslash \{\cdot\})$ qui est la loi conditionnelle du type de la dernière séquence génétique tirée d'une population pour compléter un échantillon en sachant le type de toutes les autres séquences génétiques de cet échantillon, où $\mathbf{H}_\tau \setminus \{\cdot\}$ dénote l'ensemble des séquences génétiques à l'état $\mathbf{H}_\tau$ sauf une séquence de type «·». Il s'agit du théorème de Bayes : $\pi(\cdot \mid \mathbf{H}_\tau \setminus \{\cdot\}) = \pi(\{\cdot\}, \mathbf{H}_\tau \setminus \{\cdot\})/\pi(\mathbf{H}_\tau \setminus \{\cdot\})$. Ainsi, $\pi(\{\cdot\}, \mathbf{H}_\tau \backslash \{\cdot\})$ représente la probabilité qu'un échantillon de séquences génétiques tirées aléatoirement parmi une population soit identique à l'ensemble $\{\{\cdot\}, \mathbf{H}_\tau \setminus \{\cdot\}\} = \mathbf{H}_\tau$.

Si l'événement de transition entre deux états est une coalescence entre deux séquences génétiques identiques, c'est-à-dire $\mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus C_{ii}^i$, il y a une séquence génétique de type i de moins à l'état $\mathbf{H}_{\tau+1}$, ce qui peut donc être noté par $\mathbf{H}_\tau \backslash \{i\}$.

On obtient la fraction $\frac{\pi(\mathbf{H}_{\tau+1})}{\pi(\mathbf{H}_\tau)}$ par :

$$\begin{aligned}\frac{\pi(\mathbf{H}_{\tau+1})}{\pi(\mathbf{H}_\tau)} &= \frac{\pi(\mathbf{H}_\tau \setminus \{i\})}{\pi(\{i\}, \mathbf{H}_\tau \setminus \{i\})} \\ &= \frac{\pi(\mathbf{H}_\tau \setminus \{i\})}{\pi(i \mid \mathbf{H}_\tau \setminus \{i\})\pi(\mathbf{H}_\tau \setminus \{i\})} \\ &= \frac{1}{\pi(i \mid \mathbf{H}_\tau \setminus \{i\})},\end{aligned} \tag{3.15}$$

où $\pi(i \mid \mathbf{H}_\tau \setminus \{i\})$ est la probabilité que la dernière séquence génétique tirée d'une population afin de former un échantillon qui contient déjà l'ensemble $\mathbf{H}_\tau \setminus \{i\}$ soit une séquence de type i.

De la même manière, on peut trouver l'expression de $\pi(\mathbf{H}_{\tau+1})/\pi(\mathbf{H}_\tau)$ pour tous les types d'événements de transition. Ce résultat est montré ci-dessous :

$$\frac{\pi(\mathbf{H}_{\tau+1})}{\pi(\mathbf{H}_\tau)} = \begin{cases} \frac{1}{\pi(i|\mathbf{H}_\tau\setminus\{i\})}, & \text{si } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus C_{ii}^i, \\ \frac{\pi(k|\mathbf{H}_\tau\setminus\{i,j\})}{\pi(i|\mathbf{H}_\tau\setminus\{i\})\pi(j|\mathbf{H}_\tau\setminus\{i,j\})}, & \text{si } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus C_{ij}^k, \\ \frac{\pi(j|\mathbf{H}_\tau\setminus\{i\})}{\pi(i|\mathbf{H}_\tau\setminus\{i\})}, & \text{si } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus M_i^j(m), \\ \frac{\pi(j|\mathbf{H}_\tau\setminus\{i\})\pi(k|\{\mathbf{H}_\tau,j\}\setminus\{i\})}{\pi(i|\mathbf{H}_\tau\setminus\{i\})}, & \text{si } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau \oplus R_i^{jk}(p). \end{cases}$$

3.5 Calcul de la probabilité du phénotype

Le dernier terme à calculer dans l'équation (3.5) concerne la probabilité du phénotype d'une séquence génétique d'un échantillon en connaissant la généalogie de l'échantillon et l'emplacement du TIV, c'est-à-dire $P(\Phi \mid G, c_T)$. Bien que la vraie position du TIV d'une séquence génétique soit inconnue, celle-ci peut être déduite par le phénotype de cette séquence. Il faut d'abord insérer sur toutes les séquences génétiques de l'échantillon à la position c_T un marqueur x_{c_T}. Ensuite, on extrait l'arbre partiel A_{c_T} (voir Figure 2.8) de ce marqueur depuis la généalogie G, et on insère un allèle de mutation causale dans une branche de cet arbre

Phénotype : $\phi \setminus$ Allèle : A	Porteur : 1	Non-porteur : 0
Cas : 1	cas et porteur (cp) $P(\phi = 1 \mid \mathrm{A} = 1) = f_1$	cas et non-porteur (cnp) $P(\phi = 1 \mid \mathrm{A} = 0) = f_0$
Témoin : 0	témoin et porteur (tp) $P(\phi = 0 \mid \mathrm{A} = 1) = 1 - f_1$	témoin et non-porteur (tnp) $P(\phi = 0 \mid \mathrm{A} = 0) = 1 - f_0$

Tableau 3.1: Tableau présentant les différentes catégories et les probabilités associées d'une séquence de l'échantillon après l'ajout d'une mutation causale dans une branche de la généalogie. L'allèle associé au caractère d'intérêt est noté par «A». A=0 correspond à l'allèle non-muté et A=1 correspond à l'allèle muté. Rappelons que $f_\cdot$ représente la probabilité d'être atteint par le caractère d'intérêt quand on a «$\cdot$» copie d'allèle muté.

partiel. On obtient de nouveau l'échantillon de séquences génétiques dont les statuts porteur/non-porteur de la mutation sont différents. En se basant sur les notations définies au début de ce chapitre, on résume les catégories de ces états possibles et leurs probabilités dans le tableau 3.1.

Notons par b la branche où on insère la mutation causale; alors, en utilisant les résultats du tableau 3.1, on a l'équation qui suit :

$$\begin{aligned} P(\Phi \mid G, c_T, b) &= \prod_{j=1}^{n} P(\phi_j \mid G, c_T, b) \\ &= \prod_{j=1}^{n_{cnp}} P(\phi_j = 1 \mid A_j = 0) \cdot \prod_{j=1}^{n_{tnp}} P(\phi_j = 0 \mid A_j = 0) \\ &\quad \cdot \prod_{j=1}^{n_{cp}} P(\phi_j = 1 \mid A_j = 1) \cdot \prod_{j=1}^{n_{tp}} P(\phi_j = 0 \mid A_j = 1) \\ &= f_0^{n_{cnp}} \cdot (1 - f_0)^{n_{tnp}} \cdot f_1^{n_{cp}} \cdot (1 - f_1)^{n_{tp}}, \end{aligned} \tag{3.16}$$

où $n_\cdot$ représente le nombre de séquences à la catégorie «$\cdot$» (voir les détails dans Tableau 3.1).

Selon la formule des probabilités totales, la probabilité conditionnelle $P(\Phi \mid G, c_T)$ peut être obtenue par l'équation suivante :

$$\begin{aligned} P(\Phi \mid G, c_T) &= \sum_{b \in B_{A_{c_T}}} P(\Phi \mid G, c_T, b) \cdot P(b \mid G, c_T) \\ &= \sum_{b \in B_{A_{c_T}}} f_0^{n_{cnp}} \cdot (1-f_0)^{n_{tnp}} \cdot f_1^{n_{cp}} \cdot (1-f_1)^{n_{tp}} \cdot \frac{\mid b \mid}{\sum_{w \in B_{A_{c_T}}} \mid w \mid}, \end{aligned} \tag{3.17}$$

où $B_{A_{c_T}}$ représente l'ensemble des branches de l'arbre partiel A_{c_T} et $|w|$ sont les longueurs de ces branches. La probabilité $P(b \mid G, c_T)$ est calculée à l'aide d'une propriété montrée dans le chapitre 2 : l'espérance du nombre de mutations par branche est proportionnelle à la longueur de cette branche. Donc, en sachant qu'il y a seulement une mutation apparue sur une branche d'un arbre, la probabilité qu'une branche donnée soit choisie pour insérer la mutation causale est proportionnelle à la longueur de cette branche.

Finalement, on met ensemble les équations pour obtenir la valeur théorique de la fonction de la vraisemblance :

$$\begin{aligned} L(c_T) &\approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(\Phi \mid G^{(i)}, c_T) \cdot P(G^{(i)})}{Q(G^{(i)})} \\ &= \frac{1}{M} \sum_{i=1}^{M} \left(P(\Phi \mid G^{(i)}, c_T) \cdot \prod_{\tau=0}^{\tau^*-1} \frac{P(\mathbf{H}_\tau^{(i)} \mid \mathbf{H}_{\tau+1}^{(i)})}{Q(\mathbf{H}_{\tau+1}^{(i)} \mid \mathbf{H}_\tau^{(i)})} \right) \\ &= \frac{1}{M} \sum_{i=1}^{M} \left(P(\Phi \mid G^{(i)}, c_T) \cdot \prod_{\tau=0}^{\tau^*-1} \frac{P(\mathbf{H}_\tau^{(i)} \mid \mathbf{H}_{\tau+1}^{(i)})}{P(\mathbf{H}_\tau^{(i)} \mid \mathbf{H}_{\tau+1}^{(i)}) \cdot \pi(\mathbf{H}_{\tau+1}^{(i)}) / \pi(\mathbf{H}_\tau^{(i)})} \right) \\ &= \frac{1}{M} \sum_{i=1}^{M} \left(P(\Phi \mid G^{(i)}, c_T) \cdot \prod_{\tau=0}^{\tau^*-1} \frac{\pi(\mathbf{H}_\tau^{(i)})}{\pi(\mathbf{H}_{\tau+1}^{(i)})} \right) \\ &= \frac{1}{M} \sum_{i=1}^{M} \left\{ \left(\sum_{b \in B_{A_{c_T}^{(i)}}} f_0^{n_{cnp}} \cdot (1-f_0)^{n_{tnp}} \cdot f_1^{n_{cp}} \cdot (1-f_1)^{n_{tp}} \cdot \frac{\mid b \mid}{\sum_{w \in B_{A_{c_T}^{(i)}}} \mid w \mid} \right) \right. \\ &\qquad \left. \cdot \prod_{\tau=0}^{\tau^*-1} \frac{\pi(\mathbf{H}_\tau^{(i)})}{\pi(\mathbf{H}_{\tau+1}^{(i)})} \right\} \end{aligned} \tag{3.18}$$

En pratique, on calcule la vraisemblance de toutes les positions candidates de TIV, $L(y_z)$, où $z \in \{y_1, y_2, ..., y_Z\}$. Par défaut, les positions candidates de TIV à choisir par la méthode DMap se situent au milieu de chaque intervalle formé par les marqueurs sur une séquence génétique.

3.6 Principe de la méthode DMap avec des données manquantes

La méthode DMap fonctionne aussi lorsqu'il y a de données manquantes dans notre échantillon. En fait, si on traite ces données manquantes comme du matériel non-ancestral, l'absence de ce type d'information n'a aucun effet sur la construction de la généalogie. C'est le même principe que celui montré à la section 2.3.2 : les portions inconnues des séquences parentales après un événement de recombinaison d'une séquence génétique ne sont pas problématiques dans une généalogie, car elles ne portent aucune information de notre échantillon.

Le programme DMap développé en C++ (Descary, 2012) offre une autre variante qui permet aussi d'estimer la position du TIV, en utilisant une statistique du χ^2. Le calcul de cette statistique sera présenté au chapitre 6.

CHAPITRE IV

IMPUTATION DES GÉNOTYPES DANS LES ÉTUDES D'ASSOCIATION

L'imputation des génotypes est une technique statistique permettant d'inférer les génotypes manquants à partir des haplotypes connus dans une population. Cette technique est beaucoup utilisée dans les études d'association sur le génome, elle prédit le génotype le plus probable d'un SNP avec génotype manquant. Cette technique fait partie intégrante de tests d'association entre les SNPs d'un échantillon (génotypés ou non) et un trait d'intérêt. Dans ce chapitre, on introduit la notion d'étude d'association sur le génome et on présente une méthode d'imputation particulière qu'on a utilisé dans ce mémoire.

4.1 Étude d'association pangénomique (GWAS)

L'*étude d'association pangénomique* (de l'anglais *genome-wide association study*, GWAS en version abrégée) a pour objectif de trouver des variations génétiques qui modifient le risque de développer des maladies particulières. Elle considère l'ensemble des variations génétiques sur le génome entier pour un échantillon d'individus afin de trouver s'il existe des variations qui sont reliées à un trait d'intérêt. Dans la recherche actuelle, ce trait est une maladie courante et complexe (multifactorielle) comme le cancer, le diabète ou une maladie mentale. Ce type d'étude permet de développer de meilleures stratégies pour détecter, traiter et

prévenir une maladie.

Habituellement, la GWAS se concentre sur des variations génétiques simples. Afin d'appliquer une GWAS, on utilise le plus souvent les données cas-témoins. On choisit d'abord deux groupes de participants : les cas atteints de la maladie et les témoins. Ensuite, on obtient l'ADN de chaque participant en extrayant un échantillon de sang ou en frottant un coton-tige à l'intérieur de la bouche pour récolter des cellules. À partir de leur ADN, des millions de variations génétiques sont lues afin de trouver ceux sous forme de SNPs. Autrement dit, chaque SNP représente un variant génétique. Enfin, on compare les fréquences alléliques des SNPs chez les cas par rapport aux fréquences chez les témoins. S'il existe un allèle d'un SNP qui apparaît significativement plus fréquemment chez les cas, on dit que ce variant est associé à la maladie, et ce SNP pourrait être considéré comme un bon marqueur de la région du génome humain qui serait associé à cette maladie.

Le génome humain contient plus de 10 millions de SNPs. Il est donc difficile et coûteux d'examiner chacun de ces SNPs pour déterminer s'il joue un rôle dans une maladie spécifique. Reich et Lander (2001) ont développé une hypothèse de *"Maladie Courante/Variant Courant"* (de l'anglais *Common Disease/Common Variant*, CD/CV en version abrégée). Cette théorie affirme que les maladies courantes sont reliées à des variants courants de la population. Elle suggère que de nombreuses maladies courantes sont associées à plusieurs SNPs courants, c'est-à-dire des SNPs qui ont une MAF de 5% ou plus (voir le détail à la section 1.3). Un projet international qui s'appelle HapMap a donc été conçu pour identifier les SNPs courants qui sont responsables des maladies courantes.

4.2 Le projet international HapMap

Le projet international HapMap est une collaboration entre des scientifiques venus des organisations publiques et privées de 6 pays : le Canada, la Chine, le Japon, le Nigéria, le Royaume-Uni et les États-Unis. Les données de ce projet sont mises à la disposition des chercheurs du monde entier.

4.2.1 Illustration du projet HapMap

Comme il a été mentionné dans le premier chapitre, chaque SNP représente une variation dans un seul nucléotide d'ADN sur un site du génome. Lorsqu'on regroupe plusieurs SNPs voisins sur un même chromosome, on l'appelle un bloc d'*haplotypes*, illustré dans la figure 4.1(a) ; où chaque ligne correspond au haplotype constitué d'une combinaison particulière d'allèles provenant des SNPs voisins. Le mot «HapMap» est une abréviation du mot anglais « *Haplotype Map*», le projet HapMap est donc une carte des blocs d'haplotypes, tout comme leurs emplacements dans le génome et leurs fréquences dans différentes populations à travers le monde. En effet, leur fréquence peut dépendre de la race, ethnie, etc. Un bloc d'haptolypes peut contenir un grand nombre de SNPs, mais en raison des fortes corrélations entre les SNPs dans ce bloc, quelques SNPs bien choisis suffisent à identifier ce bloc d'haplotype (Johnson *et al.*, 2001), et ces SNPs sont définis comme *tagSNPs* (voir figure 4.1(b)). Le HapMap décrit aussi les tagSNPs qui identifient les haplotypes. Les tagSNPs sont choisis en fonction d'une mesure de corrélation entre les SNPs. Cette corrélation est connue sous le nom de *déséquilibre de liaison* (de l'anglais *linkage disequilibrium*, LD en version abrégée) proposé par Lewontin et Kojima (1960), et elle indique le niveau auquel un allèle d'un SNP est corrélé avec un allèle d'un autre SNP.

Pour présenter la corrélation LD, introduisons d'abord quelques termes. Soit SNP_1

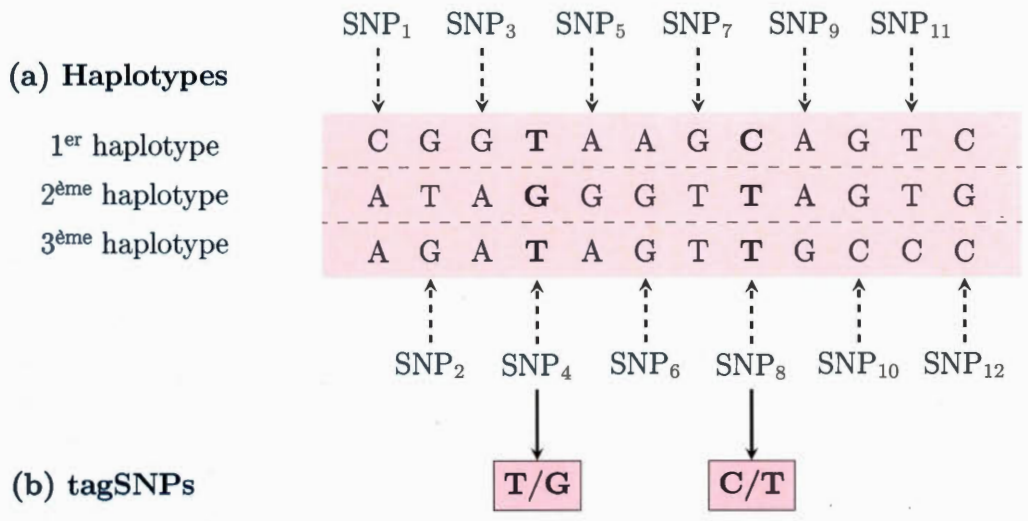


Figure 4.1: Exemple simple d'haplotypes et tagSNPs. La partie (a) représente un bloc de 3 haplotypes constitué par 12 SNPs voisins. Chaque ligne signifie un haplotype. La partie (b) montre deux tagSNPs qui sont choisis pour identifier ces haplotypes. Par exemple, si un individu a un haplotype TC sur les sites de ces deux tagSNPs, on peut estimer que son haplotype sur le site de ces 12 SNPs est le même que le 1er haplotype ici.

et SNP_2 deux SNPs possédant respectivement les allèles A/a et B/b. Dénotons par f_A et f_B les fréquences alléliques correspondantes. Ainsi, $f_a = 1 - f_A$ et $f_b = 1 - f_B$. Un individu peut avoir 3 génotypes possibles à chaque SNP : aa, Aa et AA au SNP_1 et bb, Bb et BB au SNP_2. Les combinaisons possibles des haplotypes sur les sites de ces deux SNPs sont montrés dans la figure 4.2. On remarque qu'il y a 4 haplotypes possibles : AB, Ab, aB et ab. Leur fréquence dans la population est respectivement notée par f_{AB}, f_{Ab}, f_{aB} et f_{ab}. Le tableau 4.1 illustre les relations entre les fréquences d'haplotypes et d'allèles. Une mesure du déséquilibre de liaison entre SNP_1 et SNP_2 est défini par la relation suivante :

$$D = f_{AB} \cdot f_{ab} - f_{Ab} \cdot f_{aB}. \tag{4.1}$$

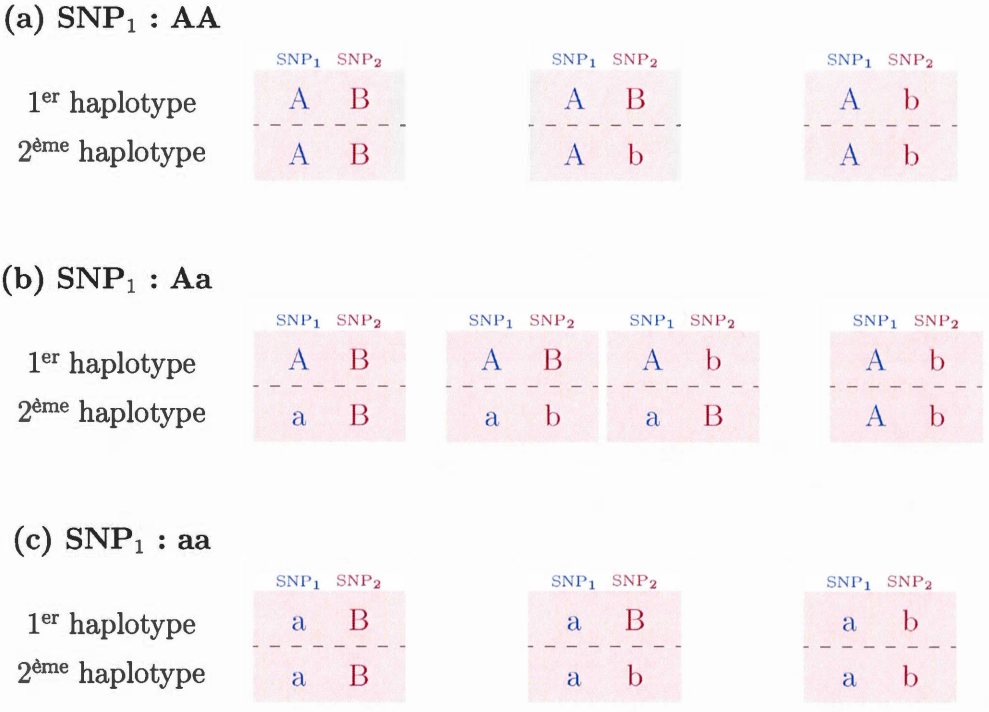


Figure 4.2: Illustration des combinaisons possibles d'haplotypes constitués de deux SNPs pour un individu. Ces deux SNPs possèdent respectivement deux allèles différents chacun : A/a et B/b. Chaque partie de cette figure nous montre des haplotypes possibles de cet individu en connaissant le génotype du SNP_1. Au total, il existe 4 haplotypes possibles : AB, Ab, aB et ab.

Cette équation peut être réécrite ainsi à l'aide du tableau 4.1,

$$D = f_{\text{AB}} - f_{\text{A}} \cdot f_{\text{B}}. \tag{4.2}$$

Ce résultat est obtenu par la propriété qui suit.

Proposition. *Supposons deux SNPs ayant les allèles A/a et B/b, un individu diploïde a quatre haplotypes possibles sur ces deux SNPs : AB, Ab, aB et ab. Soit f_A (f_a, f_B et f_b) la fréquence de l'allèle A (a, B et B) et f_{AB} (f_{Ab}, f_{aB} et f_{ab}) la fréquence de l'haplotype AB (Ab, aB et ab).*

$SNP_1 \setminus SNP_2$	B	b	Total
A	f_{AB}	f_{Ab}	f_A
a	f_{aB}	f_{ab}	f_a
Total	f_B	f_b	1

Tableau 4.1: Tableau de fréquences d'haplotypes et d'allèles.

Alors, on a :

$$(f_{AB} - f_A \cdot f_B) - (f_{AB} \cdot f_{ab} - f_{Ab} \cdot f_{aB}) = 0.$$

Démonstration :

$$\begin{aligned}
&(f_{AB} - f_A \cdot f_B) - (f_{AB} \cdot f_{ab} - f_{Ab} \cdot f_{aB}) \\
&= f_{AB}(1 - f_{ab}) - f_A \cdot f_B + f_{Ab} \cdot f_{aB} \\
&= f_{AB}(1 - (f_b - f_{Ab})) - f_A \cdot f_B + f_{Ab} \cdot f_{aB} \\
&= f_{AB} - f_b \cdot f_{AB} + f_{AB} \cdot f_{Ab} - f_A(1 - f_b) + f_{Ab} \cdot f_{aB} \\
&= f_{AB} + f_{AB} \cdot f_{Ab} - f_A + f_{Ab} \cdot f_{aB} + f_b(f_A - f_{AB}) \\
&= f_{AB} + f_{AB} \cdot f_{Ab} - f_A + f_{Ab} \cdot f_{aB} + f_b \cdot f_{Ab} \\
&= f_{AB} + f_{AB} \cdot f_{Ab} - f_A + f_{Ab} \cdot f_{aB} + f_{Ab}(1 - f_B) \\
&= f_{AB} + f_{AB} \cdot f_{Ab} - f_A - f_{Ab} \cdot f_{AB} + f_{Ab} \\
&= f_{AB} + f_{AB} \cdot f_{Ab} - (f_{AB} + f_{Ab}) - f_{Ab} \cdot f_{AB} + f_{Ab} \\
&= 0.
\end{aligned}$$

□

Clairement, l'expression (4.2) permet d'évaluer la dépendance entre les allèles A et B. Si ces deux allèles sont parfaitement indépendants, $f_{AB} = f_A \cdot f_B$, et D doit être 0. Une des mesures de LD les plus utilisées (ce qui est également appliquée par le projet HapMap) est le coefficient de corrélation, noté r^2 selon l'équation

suivante (Hill et Robertson, 1968) :

$$r^2 = \frac{D^2}{f_A \cdot f_a \cdot f_B \cdot f_b}. \tag{4.3}$$

Cette mesure reflète l'ancienneté dans l'histoire d'une population de deux SNPs, et elle varie entre 0 et 1. Si deux SNPs ont un $r^2 = 1$, cela signifie que ces deux SNPs sont parfaitement corrélés. En général, dans les études d'association, deux SNPs qui ont un r^2 de 0.8 ou plus sont considérés comme fortement corrélés. Dans ce cas, l'information de l'un de ces deux SNPs suffira pour induire celle de l'autre. Par exemple, la partie (b) à la Figure 4.1 nous montre deux tagSNPs qui permettent d'étudier l'ensemble des 3 haplotypes constitués de 12 SNPs. Ces deux tagSNPs sont choisis car les r^2 entre ceux-ci et les autres SNPs sont supérieurs ou égales à 0.8.

4.2.2 Description des données de HapMap 1, 2 & 3

D'après le Consortium International de HapMap (The International HapMap Consortium, 2005), le projet international de HapMap a été lancé en 2002 dans le but de fournir une ressource publique pour accélérer la recherche génétique médicale. Lors de la Phase I, on avait choisi 270 individus issus de 4 populations différentes :

- 30 trios (c'est-à-dire un adulte et ses deux parents), d'Ibadan au Nigéria (YRI en version abrégée, la même chose pour les autres populations) ;
- 30 trios d'habitants de l'Utah originaires du Nord et de l'ouest de l'Europe (CEU) ;
- 45 individus non reliés de Tokyo au Japon (JPT) ;
- 45 individus non reliés provenant de la principale ethnie Han de Beijing en Chine (CHB).

Ces populations ont étés prises à cause de leur diversité géographique, puisque les individus proviennent des trois principaux continents : Europe, Asie et Afrique. L'objectif du projet HapMap était de génotyper au moins un SNP courant à toutes les 5 kilobases (équivalent à 5 000 pb) du génome dans 270 individus issus de ces 4 populations diverses. On a atteint cet objectif car environ 1.3 millions de SNPs courants ont été génotypés durant cette phase.

Le HapMap 2 est la Phase II du projet HapMap. Pour augmenter la densité des SNPs dans la carte, plus de 2.1 millions de SNPs courants ont étés génotypés dans les mêmes populations que celles de la Phase I. Le HapMap 2 permet de génotyper environ un SNP par kilobase, et il contient 25-35% des SNPs courants (The International HapMap Consortium, 2007).

Grâce aux deux premières phases du projet HapMap, on a réussi à identifier certains nouveaux loci génomiques qui influencent des maladies. Mais les études étaient limitées à cause du type de SNP ($\geq$ 5% de MAF) et la diversité des populations (seulement 3 continents). Par conséquent, le HapMap 3 a recueilli un ensemble plus grand de 1 184 échantillons provenant de 11 populations. Il comprend tous les échantillons de HapMap 1 et 2 plus d'autres individus qui proviennent des mêmes populations, c'est-à-dire, CEU, CHB, JPT et YRI. En outre, le HapMap 3 a inclus 7 autres populations (The International HapMap 3 Consortium, 2010) :

— les habitants d'origine africaine du sud-ouest des États-Unis (ASW) ;

— les Chinois dans la région métropolitaine de Denver au Colorado des États-Unis (CHD) ;

— les Indiens Gujarati à Houston au Texas des États-Unis (GIH) ;

— les Luhya à Webuye du Kenya (LWK) ;

— les Maasai à Kinyawa du Kenya (MKK) ;

— les habitants d'origine mexicaine à Los Angeles au Californie des États-Unis (MXL) ;

— les habitants de Toscane en Italie (TSI).

Cette phase a donné une bonne amélioration des SNPs rares (MAF $< 0.5\%$) et les SNPs peu courants (MAF entre 0.5% et 5%). Le coefficient de corrélation r^2 moyen combiné de ces deux types de SNPs est passé de 0.60 à la Phase II à 0.76 à la Phase III (The International HapMap 3 Consortium, 2010). Le HapMap 3 nous montre aussi que les variations rares et peu courantes sont moins partagées entre les populations que les variations courantes, car ces dernières ont un r^2 moyen beaucoup plus élevé à la même phase (il est 0.946 à la Phase II et 0.961 à la Phase III). Par rapport aux SNPs courants, on n'a pas eu beaucoup de succès sur les SNPs rares et peu courants. Un autre projet qui s'appelle 1 000 Génomes a été donc créé pour continuer sur ce sujet là.

4.3 Nécessité d'imputer des génotypes

4.3.1 Génotypes manquants

Dans les études génétiques et génomiques, on vise à identifier et caractériser des variants génétiques qui ont un impact sur les traits humains, la susceptibilité à une maladie mais aussi la taille ou le poids, par exemple. Ce genre d'étude demande un examen sur une association entre les traits d'intérêt et des séquences du génome entier pour des milliers d'individus dans une étude. Mais, en réalité, on est seulement capable de mesurer des séquences d'une partie du génome pour tous les individus, car la mesure sur le génome entier dans une étude est souvent très dispendieuse. De plus, il arrive souvent que certains individus ont des allèles manquants (ainsi des génotypes manquants) à certains SNPs en raison d'un défaut expérimental. Donc, le manque des génotypes est un problème inévitable dans

les études génomiques et génétiques. La figure 4.3 illustre un exemple simple de génotypes manquants dans un échantillon qu'on étudie.

	GÉNOTYPES
1er individu	C ? C C T A A ? C G A T A A C T G ? A T C G A ?
2ème individu	A T C T T ? A G A G A T ? T T T T ? A G C C ? A
3ème individu	C T T C ? A ? T A C G A C A C C G G A G ? C A A

Figure 4.3: Exemple simple de génotypes manquants sur certains sites du génome provenant d'un échantillon de 3 individus diploïdes. Pour un individu diploïde, chaque colonne représente un génotype constitué de deux haplotypes sur un site. Les points d'interrogation représentent des allèles manquants dans l'échantillon.

4.3.2 Impact des génotypes manquants

L'effet direct des génotypes manquants est de perdre de l'information aux positions de ces derniers, ce qui influence l'exactitude des analyses comme la GWAS et la cartographie génétique. Marchini et Howie (2010) ont présenté un exemple où les variants génétiques non génotypés peuvent être très reliés au trait d'intérêt, et si l'on ignore ces variants, la signification du test d'association va beaucoup diminuer. Beaucoup d'autres recherches ont montré aussi que les génotypes manquants ont toujours un impact sur les études génomiques et génétiques. Par exemple, Howey et Cordell (2012) ont effectué une analyse sur un seul SNP avec des données simulées en utilisant différentes probabilités pour les génotypes manquants, et leurs résultats nous ont montré que la puissance pour détecter les effets génétiques

maternels (lorsqu'ils sont correctement modélisés) diminue avec l'augmentation de la proportion des génotypes manquants ; Hackett et Broadfoot (2003) ont montré par une étude de simulation que des génotypes manquants ont un effet négatif sur la proportion des cartes génétiques correctement ordonnées. Compte tenu de ces raisons, il est important de compléter des génotypes dans l'échantillon en faisant une imputation.

4.4 Méthode d'imputation : IMPUTE

La prédiction des génotypes dans une méthode d'imputation est basée sur les haplotypes connus. Ces haplotypes connus forme le *panel de référence*. Les haplotypes des HapMap 2 et 3 font partie des panels les plus utilisés. Dans cette section, on présente une des méthodes d'imputation les plus employées, IMPUTE (Marchini *et al.*, 2007; Howie *et al.*, 2009). Dans ce mémoire, toutes les données imputées sont obtenues par cette méthode.

Supposons qu'on a deux bases de données qui sont toutes constituées de L SNPs autosomiques (non sexuels) et bialléliques (2 allèles), et que les deux allèles à chaque SNP sont codés en 0 (si l'allèle est primitif) et 1 (si l'allèle est muté). La première base de données est le panel de référence qui contient K haplotypes connus des L SNPs. Désignons par $\mathbf{H} = \{\mathbf{H}_1, \ldots, \mathbf{H}_K\}$ l'ensemble de ces haplotypes, où $\mathbf{H}_k = \{H_{k1}, \ldots, H_{kL}\}$, $k = 1, \ldots, K$, représente le $k^{\text{ème}}$ haplotype des L SNPs, et $H_{kj} \in \{0, 1\}$ (voir Figure 4.4). La deuxième base de données est l'échantillon qu'on étudie, qui consiste en N individus en présence de génotypes manquants. À chaque SNP, les génotypes possibles dans cet échantillon sont 00, 10, 11 et *génotype manquant*. Dénotons G_{ij} la valeur du génotype d'un individu i au SNP_j ; cette valeur G_{ij} représente le nombre d'allèles mutés, donc $G_{ij} \in \{0, 1, 2, \text{na}\}$, où «na» représente une valeur manquante. Dans la suite de

ce mémoire, on désigne le génotype par sa valeur. L'ensemble des génotypes de l'échantillon est représenté par $\mathbf{G} = \{\mathbf{G}_1, \ldots, \mathbf{G}_N\}$, où $\mathbf{G}_i = (G_{i1}, \ldots, G_{iL})$, $i = 1, \ldots, N$, dénote le vecteur des génotypes du $i^{\text{ème}}$ individu (voir Figure 4.5).

	$\mathbf{SNP_1}$	$\mathbf{SNP_2}$	$\cdots$	$\mathbf{SNP_L}$
$\mathbf{H_1}$	H_{11}	H_{12}	$\cdots$	H_{1L}
$\mathbf{H_2}$	H_{21}	H_{22}	$\cdots$	H_{2L}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{H_K}$	H_{K1}	H_{K2}	$\cdots$	H_{KL}

Figure 4.4: Figure schématique d'un panel de référence constitué de K haplotypes connus de L SNPs. Chaque ligne représente un haplotype.

	$\mathbf{SNP_1}$	$\mathbf{SNP_2}$	$\cdots$	$\mathbf{SNP_L}$
$\mathbf{G_1}$	G_{11}	G_{12}	$\cdots$	G_{1L}
$\mathbf{G_2}$	G_{21}	G_{22}	$\cdots$	G_{2L}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{G_N}$	G_{N1}	G_{N2}	$\cdots$	G_{NL}

Figure 4.5: Figure schématique de N génotypes de L SNPs provenant de N individus. Chaque ligne représente l'ensemble des génotypes d'un individu.

Ainsi, dans la méthode IMPUTE, l'ensemble des génotypes est séparé en deux parties : les génotypes observés $\mathbf{G}_O$ et les génotypes manquants $\mathbf{G}_M$. Pour pouvoir imputer des génotypes manquants, on a besoin de trouver la distribution des génotypes manquants conditionnellement aux génotypes observés et aux

haplotypes connus. On obtient cette distribution à l'aide du théorème de Bayes :

$$\begin{aligned} P(\mathbf{G}_M \mid \mathbf{G}_O, \mathbf{H}) &= \frac{P(\mathbf{G}_M, \mathbf{G}_O, \mathbf{H})}{P(\mathbf{G}_O, \mathbf{H})} \\ &= \frac{P(\mathbf{G}_M, \mathbf{G}_O \mid \mathbf{H})P(\mathbf{H})}{P(\mathbf{G}_O \mid \mathbf{H})P(\mathbf{H})} \\ &= \frac{P(\mathbf{G}_M, \mathbf{G}_O \mid \mathbf{H})}{P(\mathbf{G}_O \mid \mathbf{H})} \\ &\propto P(\mathbf{G}_M, \mathbf{G}_O \mid \mathbf{H}). \end{aligned} \tag{4.4}$$

En supposant l'indépendance conditionnelle entre les génotypes des individus (Marchini *et al.*, 2007), la distribution conjointe des génotypes manquants et des génotypes observés peut s'écrire comme ci-dessous :

$$\begin{aligned} P(\mathbf{G}_M, \mathbf{G}_O \mid \mathbf{H}) &= P(\mathbf{G} \mid \mathbf{H}) \\ &= \prod_{i=1}^{N} P(\mathbf{G}_i \mid \mathbf{H}), \end{aligned} \tag{4.5}$$

où $P(\mathbf{G}_i \mid \mathbf{H})$ représente la distribution du vecteur de génotypes d'un individu i, sachant l'ensemble des haplotypes dans le panel de référence.

Pour un individu i, $i \in \{1, \ldots, N\}$, dont le génotype au SNP_l, $l \in \{1, \ldots, L\}$, est inconnu, l'imputation se base sur les K haplotypes du panel de référence. Ainsi, pour cet individu on définit le couple $(Z_{il}^{(1)}; Z_{il}^{(2)})$, $l \in \{1, \ldots, L\}$, où $Z_{il}^{(1)}$ et $Z_{il}^{(2)}$ indiquent la position dans le panel de référence de l'haplotype choisi. Par exemple, si $Z_{i1}^{(1)} = 2$, et $Z_{i1}^{(2)} = 4$, le génotype du $i^{\text{ème}}$ individu au SNP_1 (qui indique un couple de valeurs) correspond aux haplotypes H_{21} et H_{41} du panel de référence. Le couple $(Z_{il}^{(1)}; Z_{il}^{(2)})$ indique les haplotypes dans le panel de référence à recopier. Posons $\mathbf{Z}_i^{(j)} = \{Z_{i1}^{(j)}, \ldots, Z_{iL}^{(j)}\}$, $j \in \{1, 2\}$ et $i \in \{1, \cdots, K\}$ le vecteur d'haplotypes à recopier sur chaque site des SNPs dans le panel de référence. Cette attribution (recopiage) se fait selon (4.5). La probabilité $P(\mathbf{G}_i \mid \mathbf{H})$ est calculée

comme ci-dessous :

$$
\begin{aligned}
P(\mathbf{G}_i \mid \mathbf{H}) &= \sum_{(\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)})} P\Big(\mathbf{G}_i, (\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)}) \mid \mathbf{H}\Big) \\
&= \sum_{(\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)})} \frac{P\Big(\mathbf{G}_i, (\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)}), \mathbf{H}\Big)}{P(\mathbf{H})} \\
&= \sum_{(\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)})} \frac{P\Big(\mathbf{G}_i, (\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)}), \mathbf{H}\Big)}{P\Big((\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)}), \mathbf{H}\Big)} \cdot \frac{P\Big((\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)}), \mathbf{H}\Big)}{P(\mathbf{H})} \\
&= \sum_{(\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)})} P\Big(\mathbf{G}_i \mid (\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)}), \mathbf{H}\Big) \cdot P\Big((\mathbf{Z}_i{}^{(1)};\mathbf{Z}_i{}^{(2)}) \mid \mathbf{H}\Big), \qquad (4.6)
\end{aligned}
$$

où $\mathbf{H}$ est l'ensemble des haplotypes du panel de référence.

On doit expliciter les deux probabilités d'un terme de la dernière somme en (4.6).

L'expression (I) $= P\Big((\mathbf{Z}_i^{(1)};\mathbf{Z}_i^{(2)}) \mid \mathbf{H}\Big)$ représente la probabilité *a priori* de l'évolution des haplotypes recopiés le long de la séquence de l'individu i.

L'expression (II) $= P\Big(\mathbf{G}_i \mid (\mathbf{Z}_i^{(1)};\mathbf{Z}_i^{(2)}), \mathbf{H}\Big)$ représente la probabilité du génotype de l'individu i sachant qu'il a recopié les haplotypes $(\mathbf{Z}_i^{(1)};\mathbf{Z}_i^{(2)})$ dans le panel de référence.

Afin de trouver le terme (I), Li et Stephens (2003) proposent un modèle de Markov où la transition entre les états dépend du taux de recombinaison de la population à plusieurs intervalles (à l'échelle fine) sur le génome. Pour avoir une bonne qualité d'imputation, ces intervalles doivent couvrir les régions où il y a des génotypes manquants. L'état initial de cette chaîne de Markov est uniformément distribué sur les K^2 états (2 séquences de K états chacune), donc sa distribution s'écrit comme suit :

$$P(Z_{i1}^{(1)}, Z_{i1}^{(2)} \mid \mathbf{H}) = \frac{1}{K^2}. \qquad (4.7)$$

Dénotons $P\Big((Z_{i,l+1}^{(1)}, Z_{i,l+1}^{(2)}) \mid (Z_{il}^{(1)}, Z_{il}^{(2)}), \mathbf{H}\Big)$ la probabilité de passage du site l au

site $(l+1)$; elle se définit par les équations suivantes :

(i) $(A-B)^2$, si $Z_{il}^{(1)} = Z_{i,l+1}^{(1)}$, $Z_{il}^{(2)} = Z_{i,l+1}^{(2)}$;

(ii) $(A-B)\,B$, si $Z_{il}^{(1)} = Z_{i,l+1}^{(1)}$, $Z_{il}^{(2)} \neq Z_{i,l+1}^{(2)}$ ou $Z_{il}^{(1)} \neq Z_{i,l+1}^{(1)}$, $Z_{il}^{(2)} = Z_{i,l+1}^{(2)}$;

(iii) B^2, si $Z_{il}^{(1)} \neq Z_{i,l+1}^{(1)}$, $Z_{il}^{(2)} \neq Z_{i,l+1}^{(2)}$.

(4.8)

Dans cette équation, $A = e^{-\frac{\rho_l}{K}}$, $B = (1 - e^{-\frac{\rho_l}{K}})/K$. Le paramètre $\rho_l = 4N_e r_l$ représente le taux de recombinaison de la population entre les sites l et $l+1$, où r_l est le taux de recombinaison par génération entre les sites l et $l+1$ et N_e est la taille réelle de la population qui peut être définie par l'utilisateur du programme. Il a été montré par Marchini *et al.* (2007) que l'estimateur de N_e n'a pas d'impact sur la performance de la méthode IMPUTE.

Donc, selon la propriété d'une chaîne de Markov, la probabilité $P(\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)} \mid \mathbf{H})$ peut être écrite comme suit, en considérant seulement des dépendances entre des sites consécutifs :

$$P\left((\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}) \mid \mathbf{H}\right) = P\left((Z_{i1}^{(1)}, Z_{i1}^{(2)}) \mid \mathbf{H}\right) \prod_{l=1}^{L-1} P\left((Z_{i,l+1}^{(1)}, Z_{i,l+1}^{(2)}) \mid (Z_{il}^{(1)}, Z_{il}^{(2)}), \mathbf{H}\right)$$
$$= \frac{1}{K^2} \prod_{l=1}^{L-1} P\left((Z_{i,l+1}^{(1)}, Z_{i,l+1}^{(2)}) \mid (Z_{il}^{(1)}, Z_{il}^{(2)}), \mathbf{H}\right). \tag{4.9}$$

Revenons à l'équation (4.6) et calculons maintenant l'expression (II). Elle reflète comment les génotypes observés se rapprochent des haplotypes recopiés. Cette probabilité peut s'écrire comme ci-dessous :

$$P\left(\mathbf{G}_i \mid (\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}), \mathbf{H}\right) = \prod_{l=1}^{L} P\left(G_{il} \mid (Z_{il}^{(1)}, Z_{il}^{(2)}), \mathbf{H}\right)$$
$$= \prod_{l=1}^{L} P\left((H_{Z_{il}^{(1)},l} + H_{Z_{il}^{(2)},l}) \rightarrow G_{il}\right), \tag{4.10}$$

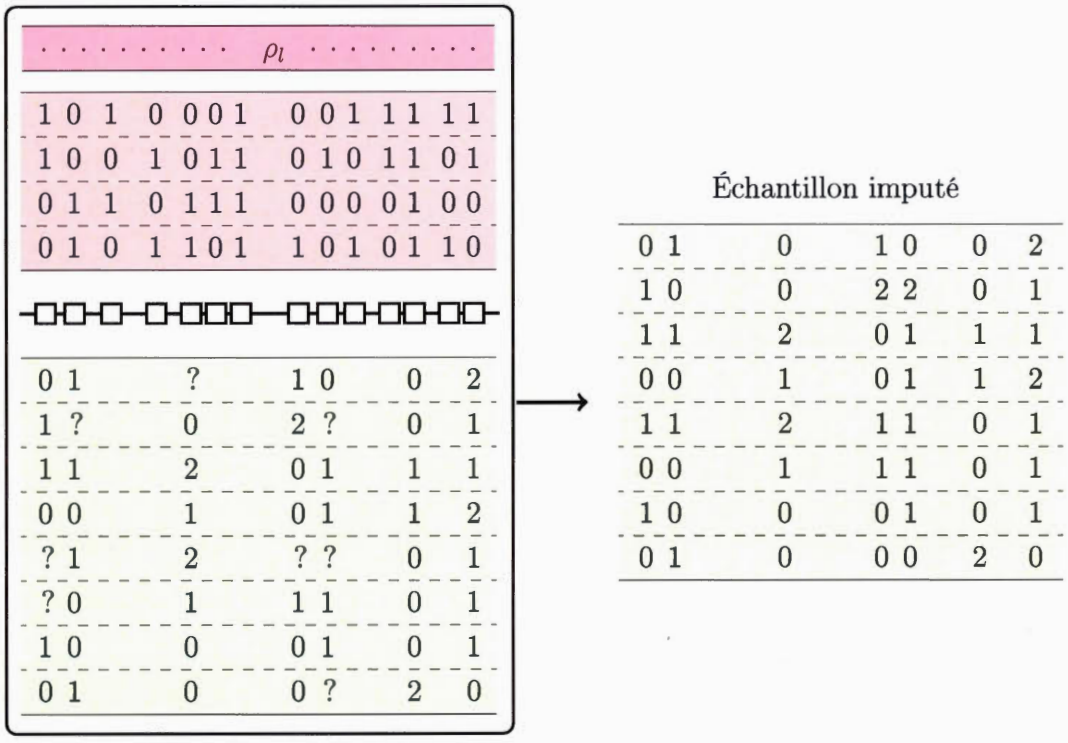
où $(H_{Z_{il}^{(1)},l} + H_{Z_{il}^{(2)},l})$ représente la somme du nombre d'allèles (à risque) recopiés depuis le panel de référence, c'est-à-dire le génotype estimé de $i^{\text{ème}}$ individu au $l^{\text{ème}}$

SNP. Un résumé du calcul de cette probabilité est donné dans le tableau 4.2 (Li et Stephens, 2003), où le paramètre λ représente la probabilité de mutation d'un allèle recopié en supposant que les mutations à tous les sites sur le génome sont indépendantes. D'après Li et Stephens (2003), cette probabilité est équivalente à $\theta/\left(2(\theta+K)\right)$, où $\theta = (\sum_{i=1}^{K-1} \frac{1}{i})^{-1}$ est le taux estimé de mutation de la population.

$\left(H_{Z_{il}^{(1)},l} + H_{Z_{il}^{(2)},l}\right) \setminus G_{il}$	0	1	2
0	$(1-\lambda)^2$	$2\lambda(1-\lambda)$	λ^2
1	$\lambda(1-\lambda)$	$\lambda^2+(1-\lambda)^2$	$\lambda(1-\lambda)$
2	λ^2	$2\lambda(1-\lambda)$	$(1-\lambda)^2$

Tableau 4.2: Illustration du résultat de la probabilité $P\left((H_{Z_{il}^{(1)},l} + H_{Z_{il}^{(2)},l}) \rightarrow G_{il}\right)$. Le paramètre $\lambda = \theta/\left(2(\theta+K)\right)$ représente la probabilité de mutation d'un allèle recopié, où $\theta = (\sum_{i=1}^{K-1} \frac{1}{i})^{-1}$.

Par conséquent, on pourrait obtenir la probabilité $Pr(\mathbf{G}_i \mid \mathbf{H})$ selon l'équation (4.6), et ensuite calculer la probabilité marginale $P(\mathbf{G}_M \mid \mathbf{G}_O, \mathbf{H})$ selon les équations (4.4) et (4.5). Pour conclure, si le génotype du $i^{\text{ème}}$ individu au $j^{\text{ème}}$ SNP est manquant, la probabilité $P(\mathbf{G}_M \mid \mathbf{G}_O, \mathbf{H})$ obtenue pour ce génotype d'après la méthode IMPUTE est sous forme d'une combinaison de trois probabilités : $P(G_{ij}=0)$, $P(G_{ij}=1)$ et $P(G_{ij}=2)$. Le génotype estimé correspond à celui qui possède la plus grande probabilité parmi les trois. Tel qu'indiqué, la figure 4.6 illustre le principe de la méthode IMPUTE. Dans le prochain chapitre, on présente une méthode qui permet de tester l'association entre des SNPs et un trait d'intérêt en utilisant des données imputées par la méthode IMPUTE.



ρ_l Taux de recombinaison de la population dans chaque intervalle

1 0 Panel de référence constitué des haplotypes (K=4)

0 1 Échantillon constitué des génotypes (N=8)

Sites des SNPs sur le génome

? Génotypes manquants dans l'échantillon

Figure 4.6: Illustration schématique du principe de la méthode IMPUTE. Chaque colonne représente un SNP sur un site spécifié du génome. Dans le panel de référence chaque ligne représente un haplotype (suite de 0 et 1) ; dans l'échantillon, chaque ligne représente un génotype (0, 1, 2 ou "?"). Rappelons que les valeurs 0, 1 et 2 du génotype correspondent respectivement à une composition d'haplotypes de 00, 10 et 11. Le génotype est manquant si un des deux allèles dans ce dernier est manquant. La figure de gauche représente toutes les informations nécessaires pour appliquer la méthode IMPUTE, la figure de droite montre l'échantillon complété après l'imputation. Cette figure est inspiré de Howie *et al.* (2009).

CHAPITRE V

TESTS D'ASSOCIATION AVEC DES DONNÉES IMPUTÉES

Afin de tester l'association entre le génotype et le phénotype avec des données imputées, on présente deux méthodes fréquentistes. Chacune de ces deux méthodes a des avantages et des inconvénients ; le choix de méthode dépend de l'information détenue par l'utilisateur. Si on veut tenir compte entièrement de l'incertitude des génotypes imputés lors de ces tests, il faut utiliser une fonction de vraisemblance en présence de données manquantes. Ce chapitre débute donc en introduisant cette approche ; ensuite on décrit deux tests classiques de type fréquentiste.

5.1 Théorie générale de la fonction de vraisemblance en présence de données manquantes

Supposons qu'on a un échantillon où il existe des données manquantes. Notons $\mathbf{Y}$ l'ensemble de nos données séparé en deux parties : données observées $\mathbf{Y}_O$ et données manquantes $\mathbf{Y}_M$; ainsi $Y = (\mathbf{Y}_O, \mathbf{Y}_M)$. D'après la théorie statistique des données manquantes (Little et Rubin, 2002), la vraisemblance correcte à utiliser dans cette situation est la log-vraisemblance des données observées qui s'écrit comme suit :

$$\ell^*(\theta) = \log P(\mathbf{Y}_O|\theta) = \log \left(\int P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)\, d\mathbf{Y}_M \right), \tag{5.1}$$

où $P(\mathbf{Y}_O, \mathbf{Y}_M|\theta) = P(\mathbf{Y}|\theta)$ est la vraisemblance des données complètes, paramétrée par un certain paramètre d'intérêt θ ; $\ell(\theta) = \log P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)$. Dans ce contexte, on suppose toujours qu'on peut sortir la dérivée par rapport à θ à l'extérieur de l'intégrale.

Par définition, le score est la dérivée de la log-vraisemblance par rapport au paramètre. Ainsi la fonction de score des données observées s'obtient à partir des équations suivantes :

$$\begin{aligned}
S^*(\theta) =& \frac{d\ell^*(\theta)}{d\theta} \\
=& \frac{dP(\mathbf{Y}_O|\theta)}{P(\mathbf{Y}_O|\theta)} \\
=& \frac{1}{\int P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)\, d\mathbf{Y}_M} \cdot \int \frac{dP(\mathbf{Y}_O, \mathbf{Y}_M|\theta)}{d\theta}\, d\mathbf{Y}_M \\
=& \frac{1}{\int P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)\, d\mathbf{Y}_M} \cdot \int \frac{1}{P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)} \cdot \frac{dP(\mathbf{Y}_O, \mathbf{Y}_M|\theta)}{d\theta} \\
& \qquad \cdot P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)\, d\mathbf{Y}_M \\
=& \int \frac{1}{P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)} \cdot \frac{dP(\mathbf{Y}_O, \mathbf{Y}_M|\theta)}{d\theta} \cdot \frac{P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)}{\int P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)\, d\mathbf{Y}_M}\, d\mathbf{Y}_M \\
=& \int \frac{d\ell(\theta)}{d\theta} \cdot P(\mathbf{Y}_M|\mathbf{Y}_O, \theta)\, d\mathbf{Y}_M \\
=& \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[S(\theta)],
\end{aligned} \tag{5.2}$$

où $S(\theta)$ est le score des données complètes, c'est-à-dire $S(\theta) = d\ell(\theta)/d\theta$.

L'information empirique $J(\theta)$ se définit à travers la dérivée seconde de la log-vraisemblance par rapport au paramètre. Évaluons d'abord l'information

empirique des données complètes :

$$\begin{aligned} J(\theta) &= -\frac{d^2\ell(\theta)}{d\theta^2} \\ &= -\frac{d}{d\theta}\left(\frac{d\ell(\theta)}{d\theta}\right) \\ &= -\frac{d}{d\theta}\left(\frac{1}{P(\mathbf{Y}|\theta)}\cdot\frac{dP(\mathbf{Y}|\theta)}{d\theta}\right) \\ &= \frac{dP(\mathbf{Y}|\theta)/d\theta}{(P(\mathbf{Y}|\theta))^2}\cdot\frac{dP(\mathbf{Y}|\theta)}{d\theta} - \frac{1}{P(\mathbf{Y}|\theta)}\cdot\frac{d^2P(\mathbf{Y}|\theta)}{d\theta^2} \\ &= \left(\frac{1}{P(\mathbf{Y}|\theta)}\cdot\frac{dP(\mathbf{Y}|\theta)}{d\theta}\right)^2 - \frac{1}{P(\mathbf{Y}|\theta)}\cdot\frac{d^2P(\mathbf{Y}|\theta)}{d\theta^2}. \end{aligned} \tag{5.3}$$

Ainsi, on obtient que :

$$\frac{1}{P(\mathbf{Y}|\theta)}\cdot\frac{d^2P(\mathbf{Y}|\theta)}{d\theta^2} = \left(\frac{1}{P(\mathbf{Y}|\theta)}\cdot\frac{dP(\mathbf{Y}|\theta)}{d\theta}\right)^2 - J(\theta). \tag{5.4}$$

L'équation (5.4) ci-dessus nous aide à trouver l'information empirique des données observées :

$$\begin{aligned} J^*(\theta) &= -\frac{d^2\ell^*(\theta)}{d\theta^2} \\ &= -\frac{d}{d\theta}\left(\frac{1}{\int P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)\,d\mathbf{Y}_M}\cdot\int\frac{dP(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}{d\theta}\,d\mathbf{Y}_M\right) \\ &= -\frac{d}{d\theta}\left(\frac{1}{\int P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)\,d\mathbf{Y}_M}\right)\cdot\int\frac{dP(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}{d\theta}\,d\mathbf{Y}_M \\ &\quad - \frac{1}{\int P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)\,d\mathbf{Y}_M}\cdot\frac{d}{d\theta}\left(\int\frac{dP(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}{d\theta}\,d\mathbf{Y}_M\right) \\ &= \frac{1}{(\int P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)\,d\mathbf{Y}_M)^2}\cdot\int\frac{dP(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}{d\theta}\,d\mathbf{Y}_M\cdot\int\frac{dP(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}{d\theta}\,d\mathbf{Y}_M \\ &\quad - \frac{1}{\int P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)\,d\mathbf{Y}_M}\cdot\int\frac{d^2P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}{d\theta^2}\,d\mathbf{Y}_M \\ &= \left(\frac{1}{\int P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)\,d\mathbf{Y}_M}\cdot\int\frac{dP(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}{d\theta}\,d\mathbf{Y}_M\right)^2 \\ &\quad - \frac{1}{\int P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)\,d\mathbf{Y}_M}\cdot\int\frac{1}{P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}\cdot\frac{d^2P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)}{d\theta^2} \\ &\qquad\qquad \cdot P(\mathbf{Y}_O,\mathbf{Y}_M|\theta)\,d\mathbf{Y}_M. \end{aligned} \tag{5.5}$$

En vertu de (5.4), $\mathbf{Y} = (\mathbf{Y}_O, \mathbf{Y}_M)$, donc le deuxième terme ci-dessus s'écrit comme

$$\begin{aligned}
&- \int \frac{1}{P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)} \cdot \frac{d^2 P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)}{d\theta^2} \cdot P(\mathbf{Y}_M|\mathbf{Y}_O, \theta)\, d\mathbf{Y}_M \\
&= - \int \left(\left(\frac{1}{P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)} \cdot \frac{dP(\mathbf{Y}_O, \mathbf{Y}_M|\theta)}{d\theta} \right)^2 - J(\theta) \right) \cdot P(\mathbf{Y}_M|\mathbf{Y}_O, \theta)\, d\mathbf{Y}_M \\
&= - \int \left(\frac{1}{P(\mathbf{Y}_O, \mathbf{Y}_M|\theta)} \cdot \frac{dP(\mathbf{Y}_O, \mathbf{Y}_M|\theta)}{d\theta} \right)^2 \cdot P(\mathbf{Y}_M|\mathbf{Y}_O, \theta)\, d\mathbf{Y}_M \\
&\quad + \int J(\theta) \cdot P(\mathbf{Y}_M|\mathbf{Y}_O, \theta)\, d\mathbf{Y}_M. \qquad (5.6)
\end{aligned}$$

Dans le cas général (vecteur de paramètres), le produit entre les scores des données s'écrit sous forme matricielle :

$$(S^*(\theta))^2 \text{ remplacé par } S^*(\theta)S^{*\intercal}(\theta), \qquad (5.7)$$

$$(S(\theta))^2 \text{ remplacé par } S(\theta)S^{\intercal}(\theta). \qquad (5.8)$$

Alors, le résultat final de l'équation (5.5) s'écrit :

$$J^*(\theta) = \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[J(\theta)] - \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[S(\theta)S^{\intercal}(\theta)] + S^*(\theta)S^{*\intercal}(\theta). \qquad (5.9)$$

Par conséquent, quand on exécute un test d'association avec des données imputées, afin de prendre en compte l'incertitude des génotypes imputés, on calcule d'abord la vraisemblance des données complètes; ensuite, on trouve la distribution $P(\mathbf{Y}_M|\mathbf{Y}_O, \theta)$ pour obtenir $S^*(\theta)$ et $J^*(\theta)$. Dans la suite, nous présentons le calcul dans le cadre des données imputées.

5.2 La vraisemblance des données complètes

Considérons qu'on détient un échantillon avec des génotypes manquants et, à l'aide de la méthode IMPUTE, les données manquantes sont complétées. Alors, nous travaillons sur l'échantillon obtenu après imputation, où cet échantillon contient N individus (constitués des SNPs). Introduisons une variable binaire qui représente

le phénotype de l'individu, et définissons par $\mathbf{\Phi} = (\Phi_1, ..., \Phi_N)$ l'ensemble des phénotypes dans l'échantillon. Alors $\Phi_i = 1$ si l'individu i est malade, ou un « cas » ; $\Phi_i = 0$ si l'individu i n'est pas malade, ou un « témoin », et supposons qu'il y a N_1 cas et N_2 témoins parmi ces N individus dans l'échantillon. Fixons un SNP et désignons par $\mathbf{G}$ l'ensemble des génotypes à ce SNP et par G_i, $i = 1, \cdots, N$ le génotype de l'individu i à ce SNP. Comme au chapitre précédent, cette dernière variable indique le nombre d'allèles mineurs (ou à risque) détenus par cet individu à ce SNP. Ainsi, la variable G_i =0, 1 ou 2 correspond respectivement au génotype AA, Aa et aa, où la lettre «A» signifie l'allèle majeur et la lettre «a» signifie l'allèle mineur. (On a omis le 2^e indice car on travaille avec un seul SNP.)

Par définition, la fonction de vraisemblance des données complètes s'obtient comme suit :

$$\begin{aligned} L(\boldsymbol{\theta}) &= P(\mathbf{Y}|\boldsymbol{\theta}) \\ &= P(\mathbf{\Phi}, \mathbf{G}|\boldsymbol{\theta}) \\ &= \frac{P(\mathbf{\Phi}|\mathbf{G}, \boldsymbol{\theta}) \cdot P(\mathbf{G}, \boldsymbol{\theta})}{P(\boldsymbol{\theta})} \\ &= P(\mathbf{\Phi}|\mathbf{G}, \boldsymbol{\theta}) \cdot P(\mathbf{G}|\boldsymbol{\theta}) \\ &\propto P(\mathbf{\Phi}|\mathbf{G}, \boldsymbol{\theta}), \end{aligned} \tag{5.10}$$

où la distribution $P(\mathbf{G}|\boldsymbol{\theta}) = P(\mathbf{G})$ ne dépend pas de $\boldsymbol{\theta}$.

Quand le phénotype est binaire, il est usuel d'utiliser un modèle de régression logistique pour modéliser une population avec des individus malades. Dans ce cas, le génotype a un effet linéaire sur le logit du risque de la maladie. Soit p_i la probabilité que l'individu i ait la maladie, alors la fonction de vraisemblance des données complètes pour un SNP s'écrit comme ci-dessous :

$$L(\boldsymbol{\theta}) \propto P(\mathbf{\Phi}|\mathbf{G}, \boldsymbol{\theta}) = \prod_{i=1}^{N} {p_i}^{\Phi_i} (1 - p_i)^{1-\Phi_i}, \tag{5.11}$$

où

$$\boldsymbol{\theta} = (\mu, \gamma), \quad \log \frac{p_i}{1-p_i} = \mu + \gamma G_i, \quad p_i = \frac{e^{\mu+\gamma G_i}}{1+e^{\mu+\gamma G_i}}. \tag{5.12}$$

Ainsi, μ et γ sont les deux paramètres qui influencent le risque de maladie dû à ce SNP. Le paramètre μ est le logit du risque de la maladie en absence de l'allèle à risque, et le paramètre γ correspond à l'augmentation du logit du risque de la maladie en ajoutant un allèle à risque. Calculons la log-vraisemblance pour ce SNP :

$$\begin{aligned}
\ell(\mu, \gamma) &= \sum_{i=1}^{N} (\Phi_i \log p_i + (1-\Phi_i) \log(1-p_i)) \\
&= \sum_{i=1}^{N} (\Phi_i (\log p_i - \log(1-p_i)) + \log(1-p_i)) \\
&= \sum_{i=1}^{N} \left(\Phi_i \log \frac{p_i}{1-p_i} + \log(1-p_i) \right) \\
&= \sum_{i=1}^{N} \left(\Phi_i (\mu + \gamma G_i) - \log(1+e^{\mu+\gamma G_i}) \right). \qquad (5.13)
\end{aligned}$$

Rappelons que lorsque le paramètre est un vecteur multidimensionnel, le vecteur score est définit par :

$$\begin{aligned}
S(\boldsymbol{\theta}) &= S(\theta_1, \cdots, \theta_d) \\
&= \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \quad \cdots \quad \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_d} \right)^{\intercal}, \qquad (5.14)
\end{aligned}$$

où d est la dimension du vecteur $\boldsymbol{\theta}$. Dans notre cas, les éléments du vecteur score s'écrivent comme ci-dessous :

$$\begin{aligned}
S_\mu &= \frac{\partial \ell(\mu, \gamma)}{\partial \mu} \\
&= \sum_{i=1}^{N} \left(\Phi_i - \frac{e^{\mu+\gamma G_i}}{1+e^{\mu+\gamma G_i}} \right) \\
&= \sum_{i=1}^{N} (\Phi_i - p_i); \qquad (5.15)
\end{aligned}$$

$$\begin{aligned} S_\gamma &= \frac{\partial \ell(\mu,\gamma)}{\partial \gamma} \\ &= \sum_{i=1}^{N} \left(\Phi_i G_i - \frac{G_i e^{\mu+\gamma G_i}}{1+e^{\mu+\gamma G_i}} \right) \\ &= \sum_{i=1}^{N} G_i(\Phi_i - p_i). \end{aligned} \tag{5.16}$$

Ainsi le vecteur score est le suivant :

$$\begin{aligned} S(\mu,\gamma) &= \begin{pmatrix} S_\mu \\ S_\gamma \end{pmatrix} \\ &= \sum_{i=1}^{N} (\Phi_i - p_i) \begin{pmatrix} 1 \\ G_i \end{pmatrix}. \end{aligned} \tag{5.17}$$

De la même manière, on obtient les éléments de la matrice de l'information empirique par les équations suivantes :

$$\begin{aligned} J_{\mu\mu} &= -\frac{\partial^2 \ell(\mu,\gamma)}{\partial \mu^2} \\ &= \sum_{i=1}^{N} \frac{e^{\mu+\gamma G_i}(1+e^{\mu+\gamma G_i}) - (e^{\mu+\gamma G_i})^2}{(1+e^{\mu+\gamma G_i})^2} \\ &= \sum_{i=1}^{N} \frac{e^{\mu+\gamma G_i}}{(1+e^{\mu+\gamma G_i})^2} \\ &= \sum_{i=1}^{N} p_i(1-p_i); \end{aligned} \tag{5.18}$$

$$\begin{aligned} J_{\gamma\gamma} &= -\frac{\partial^2 \ell(\mu,\gamma)}{\partial \gamma^2} \\ &= \sum_{i=1}^{N} G_i \frac{G_i e^{\mu+\gamma G_i}(1+e^{\mu+\gamma G_i}) - G_i(e^{\mu+\gamma G_i})^2}{(1+e^{\mu+\gamma G_i})^2} \end{aligned}$$

$$= \sum_{i=1}^{N} G_i^2 \frac{e^{\mu+\gamma G_i}}{(1+e^{\mu+\gamma G_i})^2}$$
$$= \sum_{i=1}^{N} G_i^2 p_i(1-p_i); \tag{5.19}$$

$$J_{\mu\gamma} = J_{\gamma\mu} = -\frac{\partial^2 \ell(\mu,\gamma)}{\partial\mu\partial\gamma}$$
$$= \sum_{i=1}^{N} \frac{G_i e^{\mu+\gamma G_i}(1+e^{\mu+\gamma G_i}) - G_i(e^{\mu+\gamma G_i})^2}{(1+e^{\mu+\gamma G_i})^2}$$
$$= \sum_{i=1}^{N} G_i \frac{e^{\mu+\gamma G_i}}{(1+e^{\mu+\gamma G_i})^2}$$
$$= \sum_{i=1}^{N} G_i p_i(1-p_i). \tag{5.20}$$

Finalement, on peut écrire la matrice de l'information empirique comme suit :

$$J(\mu,\gamma) = \begin{pmatrix} J_{\mu\mu} & J_{\mu\gamma} \\ J_{\gamma\mu} & J_{\gamma\gamma} \end{pmatrix}$$
$$= \sum_{i=1}^{N} p_i(1-p_i) \begin{pmatrix} 1 & G_i \\ G_i & G_i^2 \end{pmatrix}. \tag{5.21}$$

5.3 Test du multiplicateur de Lagrange

5.3.1 Contexte générique des SNPs imputées

Il a été mentionné dans le chapitre précédent que pour un génotype manquant d'un individu à un SNP, on obtient les probabilités des trois génotypes possibles selon la méthode IMPUTE. En pratique, les résultats sont présentés ainsi : chaque ligne représente un SNP, et les probabilités des trois types de génotypes AA, Aa et aa pour chaque individu sont données par trois nombres consécutifs. Un exemple

est illustré dans le tableau 5.1, ce qui montre les génotypes imputés de deux individus à un SNP provenant d'un échantillon de 26 063 individus. Désignons par $\pi_{il} = P(G_i = l|\mathbf{G_O}, \mathbf{H})$ la probabilité après l'imputation que le génotype du $i^{\text{ème}}$ individu à un SNP soit l, en conditionnant sur l'ensemble de génotypes observés $\mathbf{G_O}$ et l'ensemble d'haplotypes de référence $\mathbf{H}$. Par exemple, pour le SNP dans le tableau 5.1, $\pi_{20} = 0.65$, $\pi_{21} = 0.35$ et $\pi_{22} = 0$.

ID de SNP	Position	Allèle Majeur	Allèle Mineur	GG_1	GT_1	TT_1	GG_2	GT_2	TT_2
SNP_8	114301564	G	T	1	0	0	0.65	0.35	0

Tableau 5.1: Exemple illustrant des génotypes imputés de deux individus (parmi 26 063) d'après la méthode IMPUTE pour un SNP se situant à la position 114301564 bp d'un chromosome. Les 6 dernières colonnes représentent les génotypes de ces deux individus dont chacun est donné par 3 colonnes consécutives. Ainsi, le génotype du premier individu est GG car la probabilité d'être GT ou TT est de 0. Le génotype du deuxième individu a 65% de probabilité d'être GG et a 35% de probabilité d'être GT.

On a deux façons de procéder quand on fait le test d'association : soit on ignore l'incertitude du génotype imputé en tenant compte seulement de la plus grande probabilité, soit on prend en considération l'incertitude du génotype imputé en tenant compte des trois probabilités possibles. Dans le deuxième cas, si l'on additionne π_{il} où $l = 0, 1, 2$ sur tous les individus de l'échantillon pour le SNP du tableau 5.1, on pourrait obtenir un tableau de contingence tel que présenté dans le tableau 5.2, qui associe phénotype et génotype. Cela suggère qu'on pourrait effectuer un test du χ^2 (de façon naïve).

$\Phi \setminus \sum_{i=1}^{N} \pi_{il}$	$l=0$	$l=1$	$l=2$	Total
1	9319.85	38.15	0.00	9358
0	15613.03	90.94	1.03	15705

Tableau 5.2: Tableau de contingence entre le phénotype et le génotype pour le SNP du tableau 5.1 d'un échantillon de 26 063 individus. Ici, on tient compte de l'incertitude des génotypes imputés.

5.3.2 Principe du test

Pour tester l'association entre la maladie et le génotype à chaque SNP, on peut effectuer un test du multiplicateur de Lagrange. Selon Cox et Hinkley (1979), la statistique de ce test est donnée par l'équation suivante :

$$T = S^{\intercal}(\boldsymbol{\theta}_0) J^{-1}(\boldsymbol{\theta}_0) S(\boldsymbol{\theta}_0), \tag{5.22}$$

où $\boldsymbol{\theta}_0$ est la valeur du paramètre sous l'hypothèse nulle H_0. En général, quand la taille de l'échantillon tend vers l'infini, T suit asymptotiquement une loi de khi-deux avec d degrés de liberté, où $d = dim(\boldsymbol{\theta}_0)$.

Dans notre cas $\boldsymbol{\theta} = (\mu, \gamma)$, alors, l'hypothèse nulle revient à $\gamma = 0$ et $d = dim(\boldsymbol{\theta}_0) = 1$. L'estimateur du vecteur des paramètres est donc équivalent à $(\hat{\mu}, 0)$ où $\hat{\mu}$ est l'estimateur du maximum de vraisemblance pour μ quand $\gamma = 0$.

La log-vraisemblance sous H_0 s'obtient directement à partir de l'équation (5.13) :

$$\ell(\boldsymbol{\theta}_0) = \sum_{i=1}^{N} \left(\Phi_i \mu - \log(1 + e^{\mu}) \right). \tag{5.23}$$

On obtient $\hat{\mu}$ de la manière suivante :

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \mu} &= \frac{\partial \left(\sum_{i=1}^{N} \Phi_i \mu - \log(1+e^{\mu})\right)}{\partial \mu} \\ &= \sum_{i=1}^{N} \left(\Phi_i - \frac{e^{\mu}}{1+e^{\mu}}\right) \\ &= N_1 - \frac{N e^{\mu}}{1+e^{\mu}} = 0, \\ \Rightarrow \hat{\mu} &= \log \frac{N_1}{N_2}. \end{aligned} \tag{5.24}$$

Ainsi, selon l'équation (5.12), $\hat{p}_i = \frac{N_1/N_2}{1+N_1/N_2} = \frac{N_1}{N}$, $i = 1, 2, \cdots, N$.

À partir des fonctions de score et de l'information empirique calculées dans les équations (5.17) et (5.21), on obtient que :

$$\begin{aligned} S(\boldsymbol{\theta}_0) &= \sum_{i=1}^{N} \left(\Phi_i - \frac{N_1}{N}\right) \begin{pmatrix} 1 \\ G_i \end{pmatrix} \\ &= \sum_{i=1}^{N} \begin{pmatrix} \Phi_i - \frac{N_1}{N} & G_i\Phi_i - \frac{N_1}{N} \end{pmatrix}^{\intercal} \\ &= \begin{pmatrix} \sum_{i=1}^{N} \Phi_i - N_1 & \sum_{i=1}^{N} G_i\Phi_i - \frac{N_1}{N}\sum_{i=1}^{N} G_i \end{pmatrix}^{\intercal} \\ &= \begin{pmatrix} 0 & \sum_{i=1}^{N} G_i\Phi_i - \frac{N_1}{N}\sum_{i=1}^{N} G_i \end{pmatrix}^{\intercal}; \end{aligned} \tag{5.25}$$

$$\begin{aligned} J(\boldsymbol{\theta}_0) &= \sum_{i=1}^{N} p_i(1-p_i) \begin{pmatrix} 1 & G_i \\ G_i & G_i^2 \end{pmatrix} \\ &= \frac{N_1 N_2}{N^2} \begin{pmatrix} N & \sum_{i=1}^{N} G_i \\ \sum_{i=1}^{N} G_i & \sum_{i=1}^{N} G_i^2 \end{pmatrix}. \end{aligned} \tag{5.26}$$

Pour tenir compte de l'incertitude des génotypes imputés, on devrait appliquer la théorie développée à la section 5.1 où on écrit la fonction de vraisemblance en

présence de données manquantes. Utilisons donc $S^*(\boldsymbol{\theta}_0)$ et $J^*(\boldsymbol{\theta}_0)$ pour calculer la statistique dans l'équation (5.22). La sous-section suivante nous présente le calcul en détails.

5.3.3 Test en tenant compte de l'incertitude des données imputées

La loi des données manquantes, $P(\mathbf{Y}_M|\mathbf{Y}_O, \boldsymbol{\theta})$, est donnée par $P(G_i = l|\boldsymbol{\Phi}, \mathbf{G_O}, \mathbf{H}, \boldsymbol{\theta})$, $l = 0, 1, 2$. Rappelons que les données manquantes sont des génotypes manquants, et les données observées sont l'ensemble des phénotypes, de génotypes observés et des haplotypes de référence. Comme le phénotype et le génotype sont indépendants sous $H_0 : \gamma = 0$, et la dimension du vecteur des paramètres $\boldsymbol{\theta}$ devient 1 (c'est-à-dire $\boldsymbol{\theta}_0 = \theta$), la distribution des données manquantes $P(\mathbf{Y}_M|\mathbf{Y}_O, \theta)$, qui est égale à $P(G_i = l|\boldsymbol{\Phi}, \mathbf{G_O}, \mathbf{H}, \theta)$, devient :

$$P(G_i = l|\mathbf{G_O}, \mathbf{H}) = \pi_{il}, \quad i = 1, \cdots, N. \tag{5.27}$$

Afin de faciliter les calculs de $S^*(\boldsymbol{\theta}_0)$ et $J^*(\boldsymbol{\theta}_0)$ pour un SNP particulier, définissons les quantités suivantes :

$$e_i = \pi_{i1} + 2\pi_{i2}; \qquad f_i = \pi_{i1} + 4\pi_{i2}; \tag{5.28}$$

$$U_0 = \sum_{i:\Phi_i=0} e_i; \quad U_1 = \sum_{i:\Phi_i=1} e_i; \quad V = \sum_{i=1}^{N} f_i; \tag{5.29}$$

$$W_0 = \sum_{i:\Phi_i=0} \left(f_i - e_i^2\right); \qquad W_1 = \sum_{i:\Phi_i=1} \left(f_i - e_i^2\right). \tag{5.30}$$

Pour calculer l'espérance sous la distribution $P(\mathbf{Y}_M|\mathbf{Y}_O, \theta) = \pi_{il}$, on remarque d'abord que :

$$\sum_{i=1}^{N} \Phi_i \pi_{il} = \sum_{i:\Phi_i=1} \pi_{il}, \tag{5.31}$$

$$\sum_{i=1}^{N} \pi_{il} = \sum_{i:\Phi_i=1} \pi_{il} + \sum_{i:\Phi_i=0} \pi_{il}. \tag{5.32}$$

Ensuite, on a :

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[\sum_{i=1}^{N} G_i\Phi_i\right] &= \sum_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left(\sum_{i=1}^{N} G_i\Phi_i\right) P(\mathbf{Y}_M|\mathbf{Y}_O,\theta) \\
&= \sum_{l=0}^{2}\sum_{i=1}^{N} G_i\Phi_i\pi_{il} \\
&= \sum_{i=1}^{N}\Phi_i\pi_{i1} + 2\sum_{i=1}^{N}\Phi_i\pi_{i2} \\
&= \sum_{i:\Phi_i=1}\pi_{i1} + 2\sum_{i:\Phi_i=1}\pi_{i2} \qquad (5.33) \\
&= \sum_{i:\Phi_i=1} e_i \\
&= U_1;
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[\sum_{i=1}^{N} G_i\right] &= \sum_{l=0}^{2}\sum_{i=1}^{N} G_i\pi_{il} \\
&= \sum_{i=1}^{N}\pi_{i1} + 2\sum_{i=1}^{N}\pi_{i2} \\
&= \sum_{i:\Phi_i=0}(\pi_{i1} + 2\pi_{i2}) + \sum_{i:\Phi_i=1}(\pi_{i1} + 2\pi_{i2}) \\
&= \sum_{i:\Phi_i=0} e_i + \sum_{i:\Phi_i=1} e_i \\
&= U_0 + U_1; \qquad (5.34)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[\sum_{i=1}^{N} G_i^2\right] &= \sum_{l=0}^{2}\sum_{i=1}^{N} G_i^2\pi_{il} \\
&= \sum_{i=1}^{N}\pi_{i1} + 4\sum_{i=1}^{N}\pi_{i2}
\end{aligned}$$

$$= \sum_{i=1}^{N} f_i$$
$$= V. \tag{5.35}$$

Ainsi, on obtient $S^*(\boldsymbol{\theta}_0)$ tel que défini à l'équation (5.2) en calculant d'abord $S(\boldsymbol{\theta}_0)_{11}$ et $S(\boldsymbol{\theta}_0)_{21}$ à partir de (5.25) :

$$\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[S(\boldsymbol{\theta}_0)_{11}] = 0, \tag{5.36}$$

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[S(\boldsymbol{\theta}_0)_{21}\right] &= \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[\sum_{i=1}^{N} G_i\Phi_i - \frac{N_1}{N}\sum_{i=1}^{N} G_i\right] \\ &= U_1 - \frac{N_1}{N}(U_0 + U_1) \\ &= \frac{N_2}{N}U_1 - \frac{N_1}{N}U_0. \end{aligned} \tag{5.37}$$

Cela donne :

$$\begin{aligned} S^*(\boldsymbol{\theta}_0) &= \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[S(\boldsymbol{\theta}_0)] \\ &= \begin{pmatrix} 0 & \frac{N_2}{N}U_1 - \frac{N_1}{N}U_0 \end{pmatrix}^{\intercal}. \end{aligned} \tag{5.38}$$

Selon la définition de $J^*(\boldsymbol{\theta}_0)$ dans l'équation (5.9), évaluons d'abord $\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[J(\boldsymbol{\theta}_0)]$ et $\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[S(\boldsymbol{\theta}_0)S^{\intercal}(\boldsymbol{\theta}_0)] - S^*(\boldsymbol{\theta}_0)S^{*\intercal}(\boldsymbol{\theta}_0)$. En vertu de l'équation (5.26), on a :

$$\begin{aligned} &\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[J(\boldsymbol{\theta}_0)] \\ =&\frac{N_1N_2}{N^2}\begin{pmatrix} \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[N] & \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[\sum_{i=1}^{N} G_i\right] \\ \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[\sum_{i=1}^{N} G_i\right] & \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[\sum_{i=1}^{N} G_i^2\right] \end{pmatrix} \\ =&\frac{N_1N_2}{N^2}\begin{pmatrix} N & U_0 + U_1 \\ U_0 + U_1 & V \end{pmatrix}. \end{aligned} \tag{5.39}$$

Ensuite, les équations (5.25) et (5.38) donnent :

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[S(\boldsymbol{\theta}_0)S^{\intercal}(\boldsymbol{\theta}_0)] - S^*(\boldsymbol{\theta}_0)S^{*\intercal}(\boldsymbol{\theta}_0) \\
=&\begin{pmatrix} 0 & 0 \\ 0 & \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}\left[(S(\boldsymbol{\theta}_0)_{21})^2\right] \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & (S^*(\boldsymbol{\theta}_0)_{21})^2 \end{pmatrix} \\
=&\frac{1}{N}\begin{pmatrix} 0 & 0 \\ 0 & N_1^2W_0 + N_2^2W_1 \end{pmatrix}.
\end{aligned} \tag{5.40}$$

Finalement, on obtient :

$$\begin{aligned}
J^*(\boldsymbol{\theta}_0) &= \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[J(\boldsymbol{\theta}_0)] - \mathbb{E}_{\mathbf{Y}_M|\mathbf{Y}_O,\theta}[S(\boldsymbol{\theta}_0)S^{\intercal}(\boldsymbol{\theta}_0)] + S^*(\boldsymbol{\theta}_0)S^{*\intercal}(\boldsymbol{\theta}_0) \\
&= \frac{N_1N_2}{N^2}\begin{pmatrix} N & U_0+U_1 \\ U_0+U_1 & V \end{pmatrix} - \frac{1}{N}\begin{pmatrix} 0 & 0 \\ 0 & N_1^2W_0 + N_2^2W_1 \end{pmatrix}
\end{aligned} \tag{5.41}$$

Ainsi, on peut écrire la valeur de la statistique du test du multiplicateur de Lagrange sous $H_0 : \gamma = 0$. Cette statistique revient à $T = \frac{(S_\gamma^*)^2}{J_\gamma^*}$ (Cox et Hinkley, 1979), où

$$\begin{aligned}
S_\gamma^* &= S_\gamma^*(\boldsymbol{\theta}_0) \\
&= \frac{N_2U_1 - N_1U_0}{N},
\end{aligned} \tag{5.42}$$

et

$$\begin{aligned}
J_\gamma^* &= J^*(\boldsymbol{\theta}_0)_{\gamma\gamma} - J^*(\boldsymbol{\theta}_0)_{\gamma\mu}J^{*-1}(\boldsymbol{\theta}_0)_{\mu\mu}J^*(\boldsymbol{\theta}_0)_{\mu\gamma} \\
&= \frac{N_1N_2}{N^2}\cdot V - \frac{N_1^2W_0 + N_2^2W_1}{N} - \frac{N_1N_2}{N^2}\cdot\frac{(U_0+U_1)^2}{N} \\
&= \frac{N_1N_2}{N^2}\left(V - \frac{N(N_1^2W_0 + N_2^2W_1)}{N_1N_2} - \frac{(U_0+U_1)^2}{N}\right).
\end{aligned} \tag{5.43}$$

Ainsi, la statistique T se calcule comme suit :

$$T = \frac{(N_2U_1 - N_1U_0)^2}{N_1N_2\left(V - \frac{N(N_1^2W_0+N_2^2W_1)}{N_1N_2} - \frac{(U_0+U_1)^2}{N}\right)}. \tag{5.44}$$

Sachant que $T \sim \chi^2_1$ sous H_0 quand $N \to \infty$, on peut donc effectuer un test sur le paramètre γ.

En général, ce test ne fonctionne pas bien dans les situations suivantes :

(i) la taille d'échantillon est petite ;

(ii) la fréquence de l'allèle mineur est faible ;

(iii) l'incertitude sur les génotypes imputés est grande.

Dans ce cas, le test du multiplicateur de Lagrange donne à tort de petite valeur-p, et on préfère utiliser le test du rapport de vraisemblance présenté à la section suivante (Marchini et Howie, 2010).

5.4 Test du rapport de vraisemblance

Une approche alternative au test du multiplicateur de Lagrange est de maximiser la vraisemblance directement sous H_0. La maximisation peut se faire avec l'algorithme de Newton-Raphson (Garthwaite *et al.*, 1995) qui revient à résoudre :

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + [J^*(\boldsymbol{\theta}^t)]^{-1} S^*(\boldsymbol{\theta}^t), \qquad t = 0, 1, \ldots \tag{5.45}$$

C'est une méthode itérative qui permet de trouver le maximum à l'étape $t+1$ en se basant sur la valeur du paramètre à l'étape t. En choisissant une bonne valeur de départ pour le paramètre, $\boldsymbol{\theta}^{t+1}$ converge vers $\hat{\boldsymbol{\theta}}$ qui maximise la vraisemblance. Ensuite on effectue un test du rapport de vraisemblance pour tester les hypothèses qui sont, dans notre cas :

$$H_0 : \gamma = 0 \qquad \text{contre} \qquad H_1 : \gamma \neq 0. \tag{5.46}$$

La statistique de ce test est la suivante :

$$T_{TRV} = 2\left(\ell^*(\hat{\boldsymbol{\theta}}) - \ell^*(\hat{\boldsymbol{\theta}}_0)\right), \tag{5.47}$$

où $\hat{\boldsymbol{\theta}}_0$ est l'estimateur de maximum de vraisemblance sous H_0. La loi de cette statistique est approximativement une loi du khi-deux avec 1 degré de liberté quand $N_1, N_2 \to \infty$.

Les deux méthodes fréquentistes présentées ci-dessus sont implantées dans le programme SNPTEST. Dans le prochain chapitre, on présente les résultats des tests d'association entre les SNPs imputés et les phénotypes en utilisant ce programme.

CHAPITRE VI

COMPARAISON DE MÉTHODES : ÉTUDE DE SIMULATION

Dans les chapitres 3 et 5, on a décrit deux types de méthodes qui permettent de tester le lien entre un caractère génétique et des marqueurs en présence de données manquantes. Dans ce chapitre, nous allons comparer la performance de ces méthodes à travers des données simulées. Nous allons d'abord décrire comment simuler des données qui satisfont les suppositions de notre modèle.

6.1 Simulation des données

Un grand avantage à utiliser des données simulées est qu'on connaît la position exacte du TIV dans nos échantillons. Cela nous permet d'examiner si les résultats obtenus par les méthodes sont bonnes. Dans notre cas, les données simulées sont obtenues à partir des programmes FastSimCoal (Excoffier et Foll, 2011; Excoffier *et al.*, 2013) et SimHGD (Larribe et Dupont, 2016). Afin de créer des échantillons de séquences génétiques qui contiennent des données manquantes, il y a trois étapes principales. Premièrement, le programme SimHGD génère des données par un simulateur du processus de coalescence avec recombinaison et convertit ces données à une population diploïde qui contient des cas et des témoins comme dans une étude génétique humaine. Le simulateur qu'on utilise dans ce mémoire s'appelle FastSimCoal. Il permet de simuler une population et/ou un échantillon

selon la théorie de la coalescence avec recombinaison dans laquelle un ARG est d'abord généré, puis les séquences génétiques sont présentées sous forme de SNPs. Ensuite, le programme SimHGD crée des échantillons aléatoires ou cas-témoins selon le besoin d'utilisateur ; dans ce mémoire, on choisit cas-témoins. Finalement, on enlève certaines données de nos échantillons d'une manière aléatoire dans une proportion spécifiée. En procédant ainsi, on crée des échantillons qui contiennent des génotypes à différents niveaux de proportions manquantes. Ensuite, on analyse ces données avec les méthodes DMap et SNPTEST, respectivement. On peut donc étudier l'impact de l'information manquante et de l'imputation sur ces deux types de méthodes différentes.

Afin d'analyser les résultats avec les méthodes DMap et SNPTEST, on simule d'abord une population par le programme FastSimCoal en considérant les paramètres suivants :

(i) la taille de la population : 20 000 ;

(ii) la longueur de séquence génétique : 1 Mb ;

(iii) le taux de mutation : 2×10^{-8} par bp par génération ;

(iv) le taux de recombinaison : 1×10^{-7} par bp par génération.

Ensuite, on génère 7 échantillons avec le programme SimHGD selon sept différents scénarios (voir Tableau 6.1).

Présentons maintenant l'algorithme principal du programme SimHGD. Les étapes sont les suivantes :

— calculer la fréquence de l'allèle mineur (MAF) à chaque marqueur ;

— parmi les 80% marqueurs se situant au milieu de séquences génétiques, sélectionner le marqueur dont la MAF est la plus proche du taux de mutation causal, ce marqueur sera le TIV ;

— garder seulement les marqueurs qui ont une MAF de 0.5% ou plus ;

Scénario	t_m	$F = (f_0, f_1, f_2)$	f_1/f_0	f_2/f_0	n_c	n_t
A	0.1	$F = (0.1, 0.5, 0.9)$	5	9	600	600
B_1	0.1	$F = (0.1, 0.3, 0.7)$	3	7	600	600
B_2	0.2	$F = (0.1, 0.3, 0.7)$	3	7	600	600
C_1	0.1	$F = (0.1, 0.25, 0.5)$	2.5	5	600	600
C_2	0.3	$F = (0.1, 0.25, 0.5)$	2.5	5	600	600
D_1	0.5	$F = (0.1, 0.1, 0.25)$	1	2.5	600	600
D_2	0.7	$F = (0.1, 0.1, 0.25)$	1	2.5	600	600

Tableau 6.1: Tableau illustrant les caractéristiques des échantillons générés selon des scénarios différents. t_m représente la fréquence de la maladie dans la population ; n_c et n_t correspondent au nombre de cas et de témoins dans l'échantillon, respectivement.

— regrouper les séquences 2 à 2 pour simuler les diploïdes et assigner le phénotype (cas ou témoin) à chaque séquence génétique (diploïde) selon son nombre d'allèles mutés du TIV et la fonction de pénétrance F ;

— pour chaque séquence génétique (diploïde), retirer le marqueur qui représente le TIV et garder en mémoire son emplacement ;

— sélectionner aléatoirement n_c séquences (diploïdes) de statut cas et n_t séquences (diploïdes) de statut témoin.

Ces échantillons ont la même taille et différentes fonction de pénétrance, fréquence de maladie dans la population et proportion de données manquantes. Comme les données produites par SimHGD sont sous forme de population diploïde, la fonction de pénétrance F contient 3 valeurs : f_0, f_1, f_2. Rappelons que, par exemple, f_2 représente la probabilité d'être malade lorsqu'on a 2 copies d'allèles mutés. La fraction f_i/f_0 où $i = 1, 2$ représente le risque relatif qu'une séquence d'un cas porte i copies d'allèle du TIV par rapport au fait qu'elle n'est pas

porteuse du TIV. Évidemment, plus cette valeur est grande, plus la mutation a un effet sur le caractère d'intérêt, et les deux méthodes devraient donner une meilleure estimation quand le risque relatif est élevé. Chaque échantillon généré par SimHGD contient 600 cas et 600 témoins ainsi que 539 marqueurs (sans compter le marqueur représentant le TIV), et on l'appelle échantillon d'origine.

Ensuite, on crée un échantillon de sujets à étudier et un panel de référence. En pratique, le nombre d'individus dans le panel de référence est beaucoup moins important que le nombre de sujets dans l'échantillon observé ; par contre, le nombre de marqueurs est plus grand dans le panel. Supposons qu'environ 80% de marqueurs de l'échantillon se retrouve dans le panel de référence. Alors, afin de créer un tel échantillon de sujets et son panel de référence, on suit les étapes suivantes :

Étape 1 : Choisir aléatoirement 350 marqueurs de l'échantillon d'origine, mettre les 189 restants de côté.

Étape 2 : Choisir aléatoirement 590 cas et 590 témoins de l'échantillon obtenu à l'étape précédente, ce sont nos sujets à l'étude (avec les deux méthodes).

Étape 3 : Parmi les 350 marqueurs choisis à l'étape 1, sélectionner aléatoirement 280 (environ 80%) marqueurs, combiner ces derniers avec les 189 marqueurs restants.

Étape 4 : Pour les 10 cas et 10 témoins enlevés à l'étape 2 et qui ne font pas partie des sujets à l'étude, on garde tous les marqueurs retenus à l'étape 3 (au total 280+189 marqueurs). Ces 20 sujets forment le panel de référence qui sert plus tard à imputer les génotypes manquants (seulement pour la méthode SNPTEST). Notons que le panel de référence ne contient pas d'individus de l'échantillon à étudier.

Afin qu'on puisse étudier l'impact des données manquantes (ou imputées) sur la méthode DMap (ou SNPTEST), choisissons 5 niveaux différents de proportions de données manquantes : 5%, 15%, 30%, 45% et 50%. Ces derniers représentent les pourcentages de génotypes manquants dans l'échantillon. Rappelons qu'un génotype est dit manquant si un des deux allèles dans ce dernier est manquant. Par la suite, on enlève aléatoirement certaines données de l'échantillon d'étude selon ces différentes proportions. Comme il a été mentionné aux chapitres précédents, quand il manque de données dans l'échantillon d'étude, la méthode DMap fonctionne toujours, tandis que la méthode SNPTEST fait appel à une imputation avant d'effectuer des analyses. Ainsi, on teste directement les données incomplètes avec la méthode DMap, et on teste les données imputées (par la méthode IMPUTE) avec la méthode SNPTEST. Dans les sections suivantes, on présente les résultats obtenus avec ces deux types de méthodes en utilisant des échantillons avec différents taux de données manquantes ou imputées.

6.2 Résultats obtenus avec la méthode SNPTEST

Dans cette sous-section, nous illustrons les résultats obtenus avec la méthode SNPTEST selon les différents scénarios du tableau 6.1. Sur chaque figure, on présente des graphiques représentant les résultats différents par rapport au taux de données imputées. L'axe des x représente des positions le long d'une séquence génétique de notre échantillon ; l'axe des y représente $-\log(\text{valeur-p})$. De plus, la ligne rouge verticale représente la vraie position du TIV ; le triangle bleu montre la position estimée par la méthode SNPTEST.

En examinant les résultats obtenus, on s'aperçoit que la méthode SNPTEST a une performance excellente en utilisant des données imputées quand la proportion de données imputées dans l'échantillon est moins de 45%. Au-delà de 45%, dans la

plupart des scénarios, l'imputation ne réussit pas à fournir une bonne estimation. De plus, lorsque le risque relatif est extrêmement faible et la fréquence de la maladie n'est pas très grande (Scénario D_1 : $t_m = 0.5$, $f_1/f_0 = 1$ et $f_2/f_0 = 2.5$), cette méthode a tendance à être moins efficace, car les points sur tous les graphiques sont très dispersés. Par contre, si la fréquence de la maladie devient extrêmement grande dans cette dernière situation (Scénario D_2 : $t_m = 0.7$), la méthode redevient efficace.

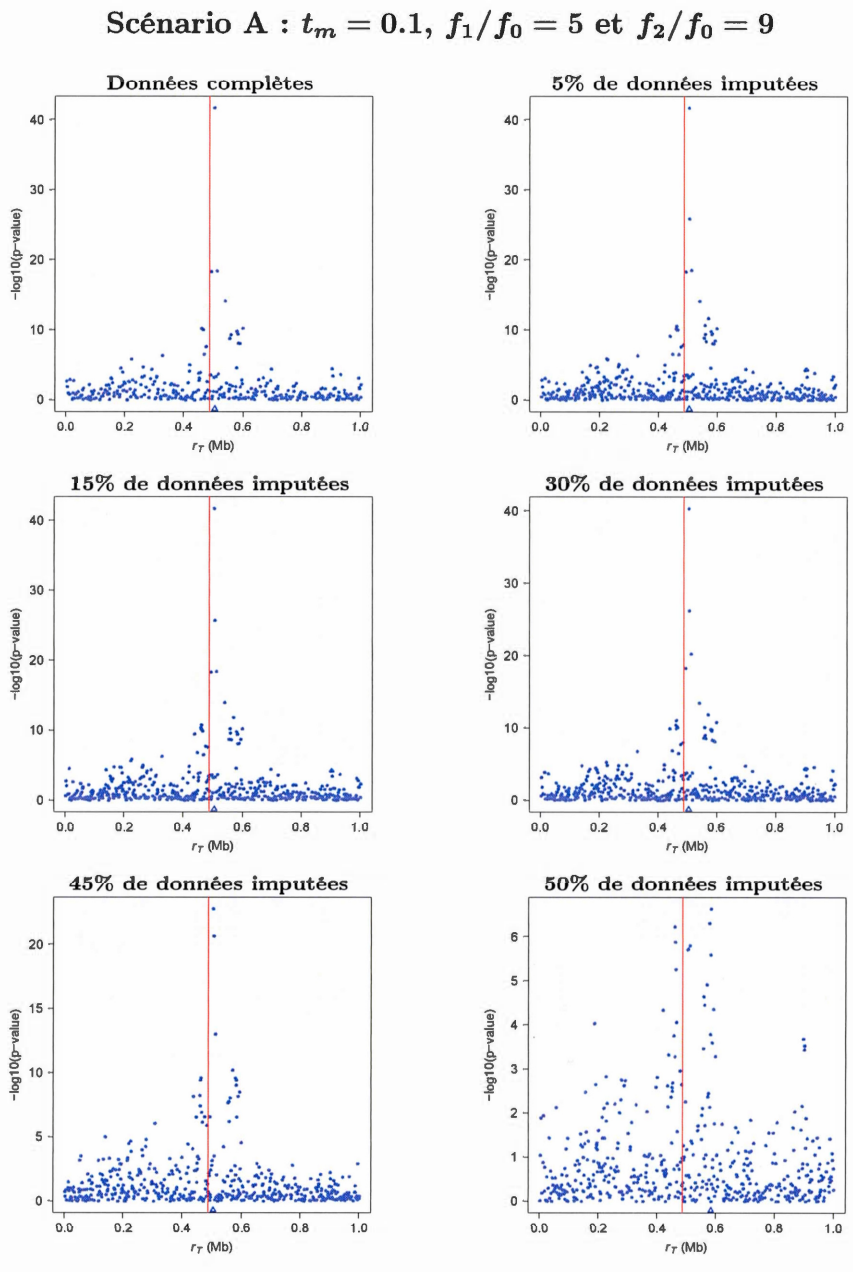


Figure 6.1: Illustration schématique des résultats de SNPTEST avec 5 niveaux différents d'imputation. L'échantillon est généré selon le scénario A.

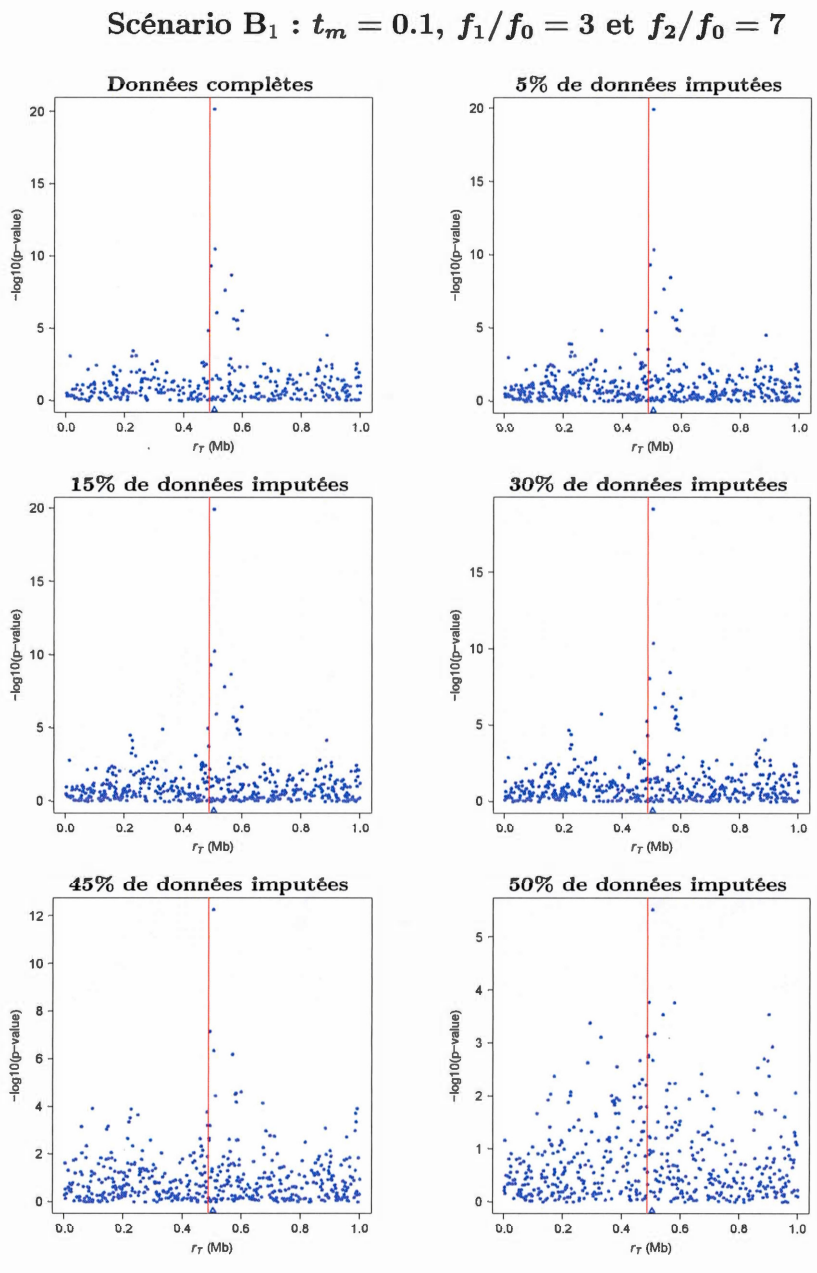


Figure 6.2: Illustration schématique des résultats de SNPTEST avec 5 niveaux différents d'imputation. L'échantillon est généré selon le scénario B_1.

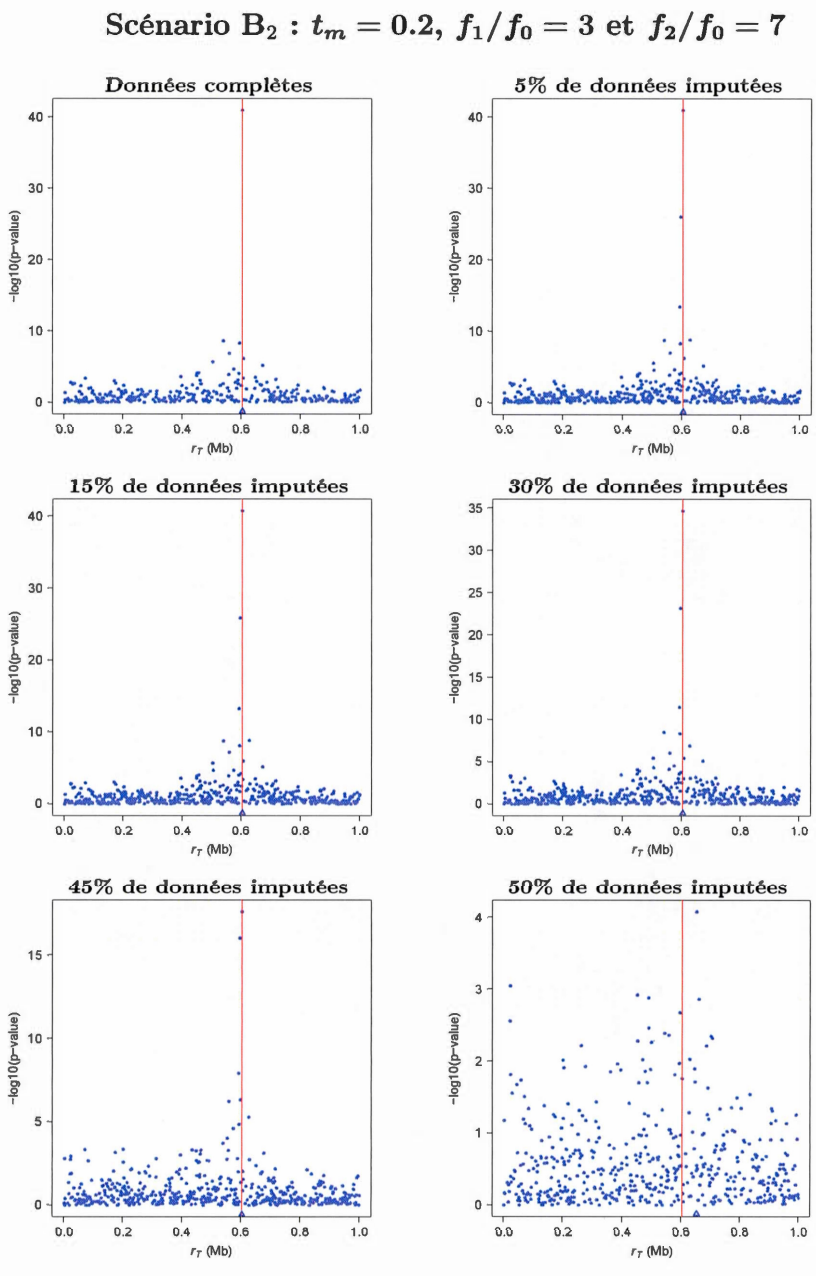


Figure 6.3: Illustration schématique des résultats de SNPTEST avec 5 niveaux différents d'imputation. L'échantillon est généré selon le scénario B_2.

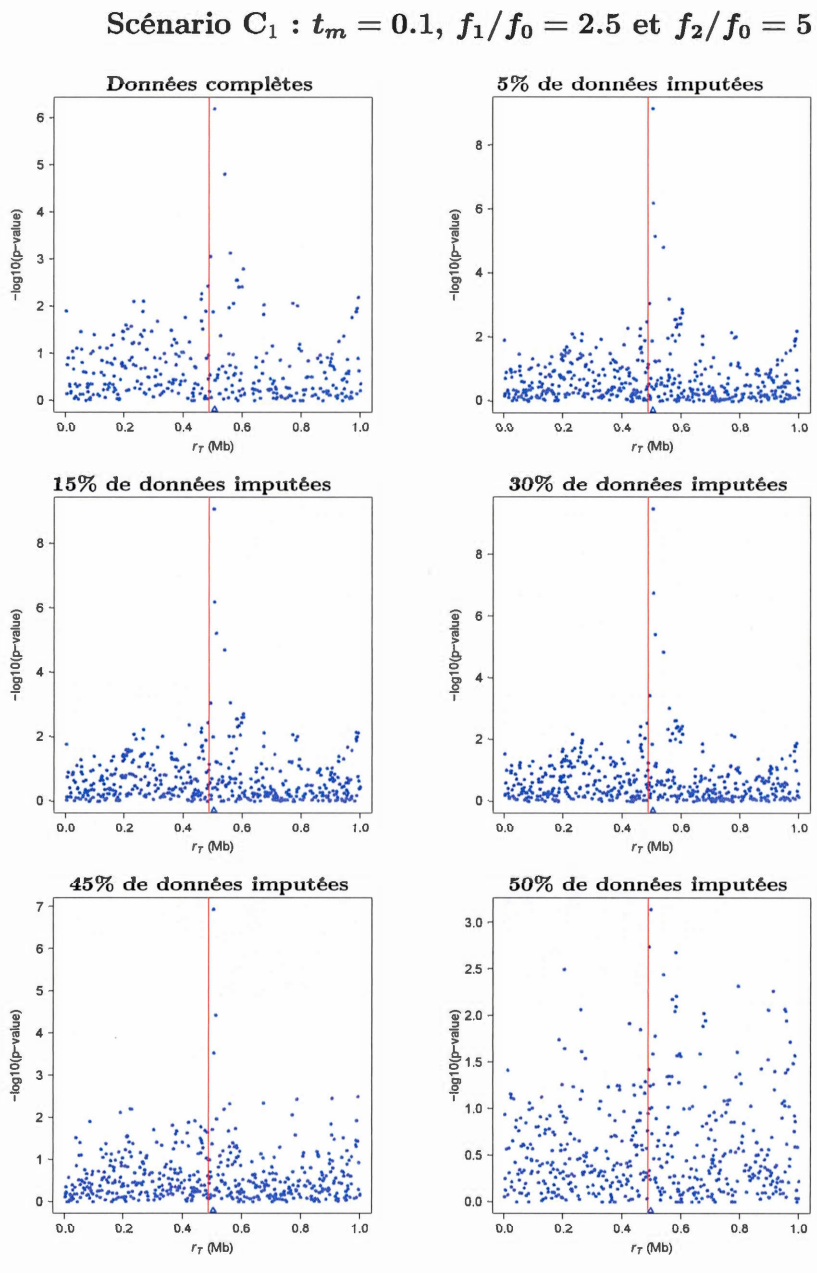


Figure 6.4: Illustration schématique des résultats de SNPTEST avec 5 niveaux différents d'imputation. L'échantillon est généré selon le scénario C_1.

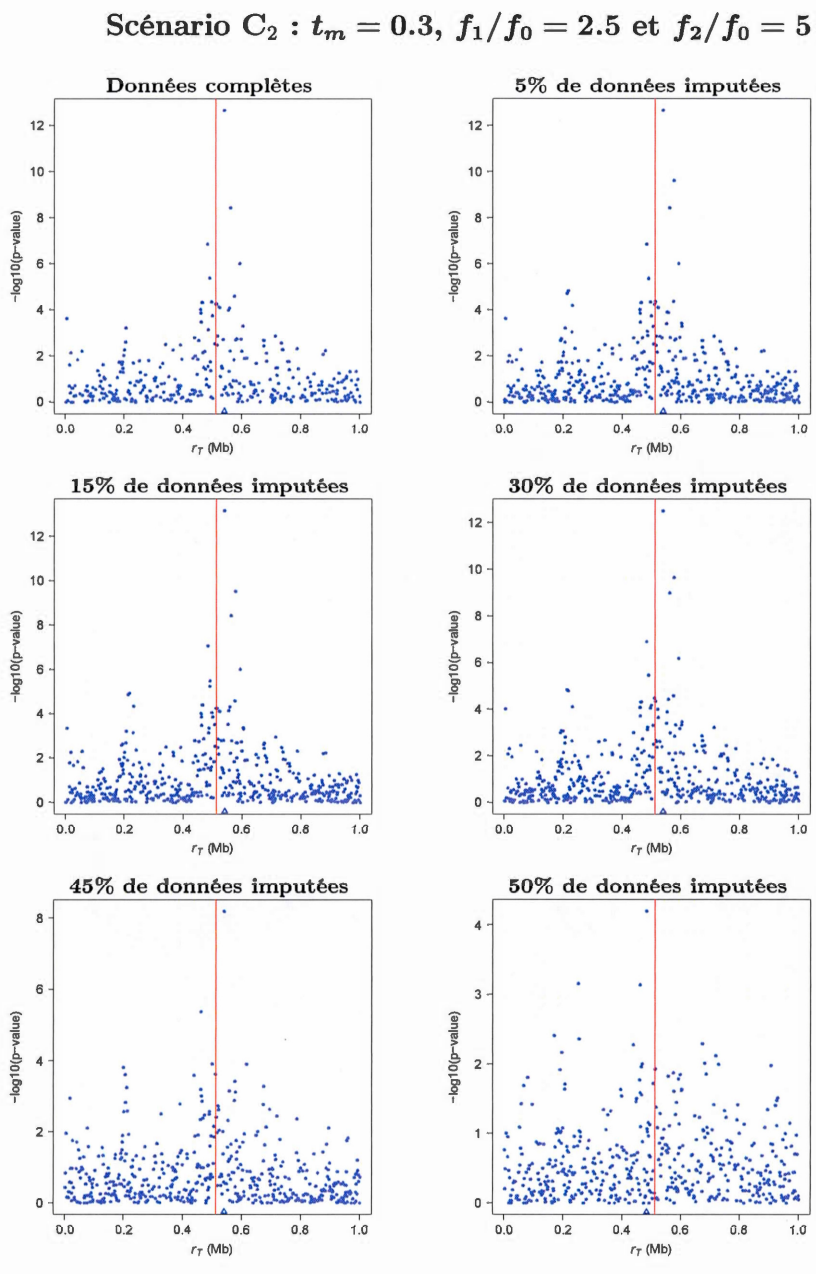


Figure 6.5: Illustration schématique des résultats de SNPTEST avec 5 niveaux différents d'imputation. L'échantillon est généré selon le scénario C_2.

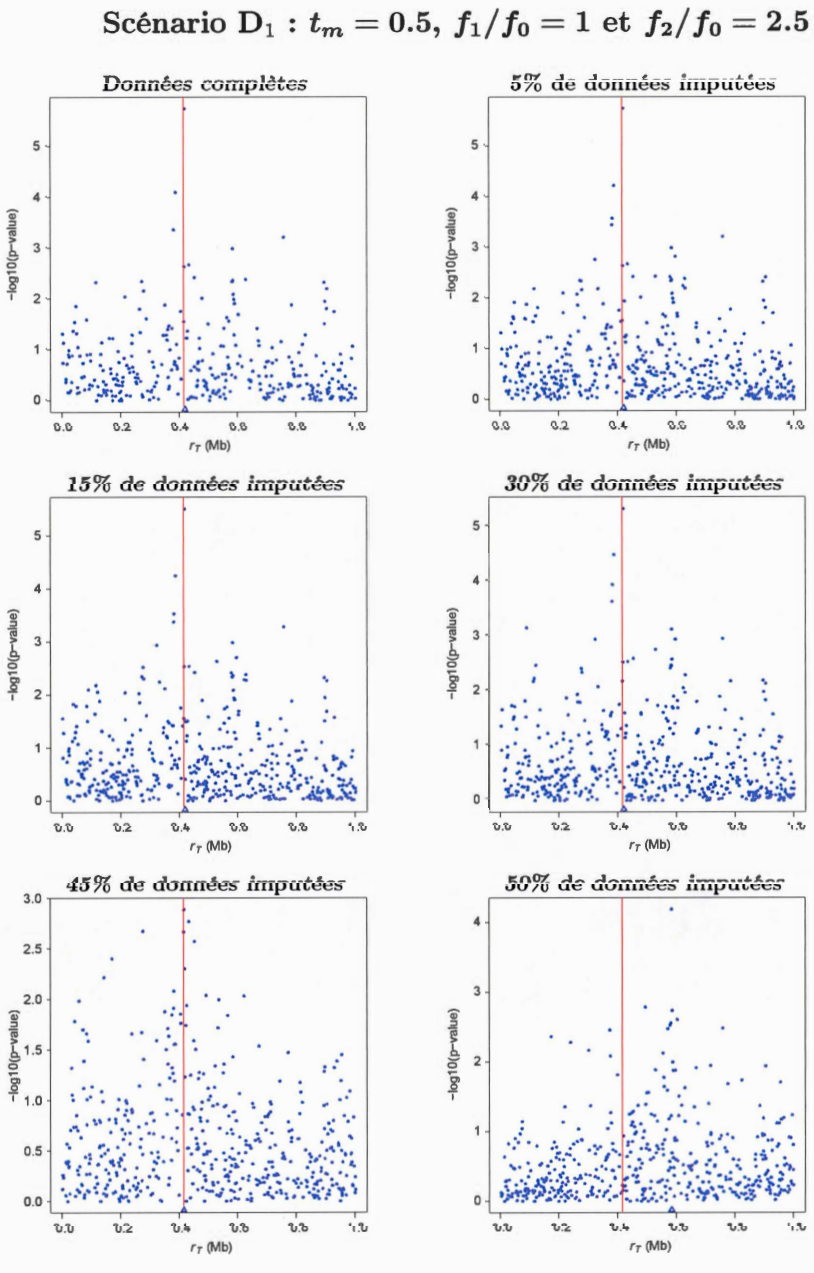


Figure 6.6: Illustration schématique des résultats de SNPTEST avec 5 niveaux différents d'imputation. L'échantillon est généré selon le scénario D_1.

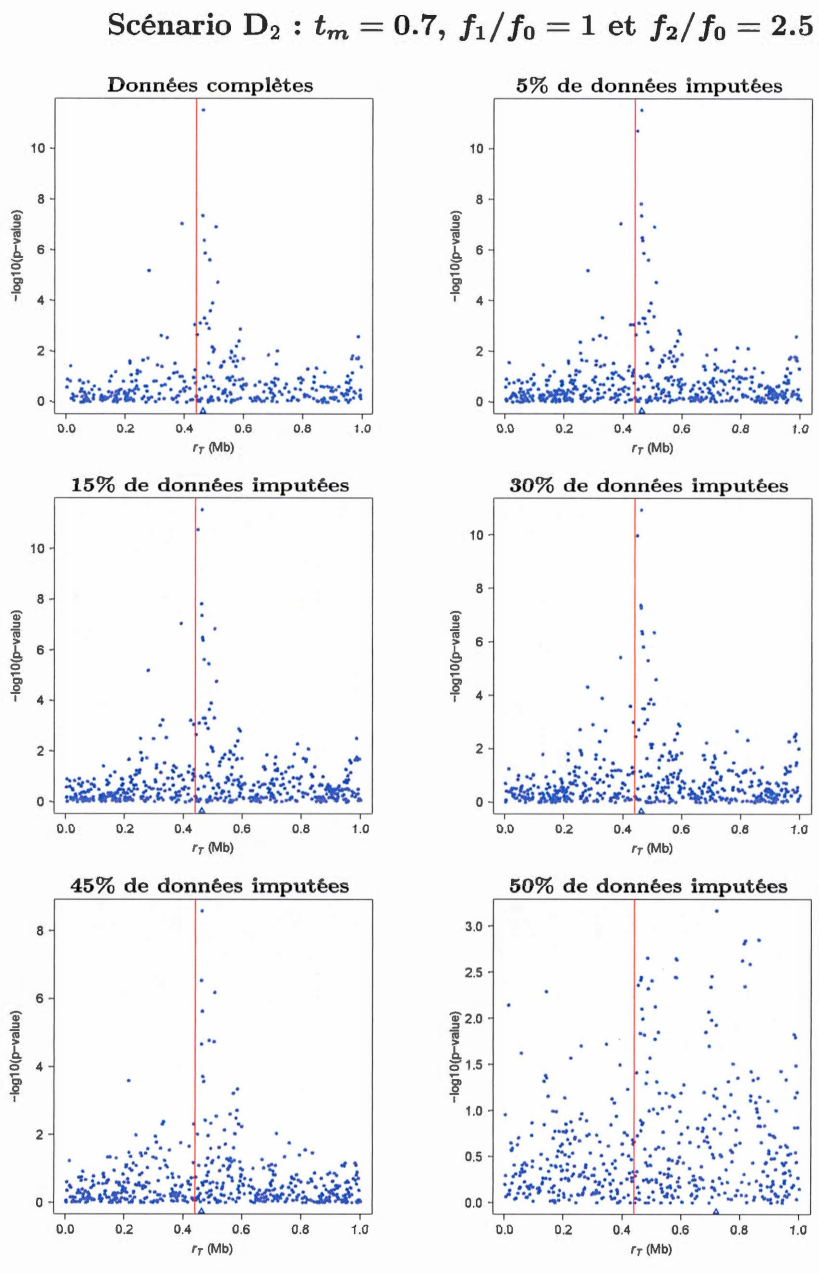


Figure 6.7: Illustration schématique des résultats de SNPTEST avec 5 niveaux différents d'imputation. L'échantillon est généré selon le scénario D_2.

6.3 Résultats obtenus avec la méthode DMap

6.3.1 Les paramètres utilisés avec la méthode DMap

Nous présentons d'abord les paramètres utilisés dans la méthode DMap. Pour trouver la position inconnue c_T du TIV, on choisit d'analyser $L = 170$ marqueurs équidistants ; cette valeur de L nous permet de maximiser la performance du programme tout en l'exécutant dans un temps raisonnable. Comme mentionné au chapitre 3, les positions candidates du TIV sont les milieux de chaque intervalle formé par les marqueurs choisis, c'est-à-dire $\{y_1, y_2, \cdots, y_{169}\}$. Le programme DMap offre une statistique de type χ^2 qui permet aussi d'estimer la position du TIV. Cette statistique demeure plus stable par rapport à la valeur de la fonction de vraisemblance quand le risque relatif est très faible. En sachant que dans la plupart de nos scénarios, les échantillons ont un risque relatif assez faible, nous utilisons donc cette statistique pour analyser nos données. Le calcul de cette statistique pour chaque position candidate suit les étapes suivantes :

- construire M ARGs avec des haplotypes. Pour chaque ARG, extraire l'arbre partiel correspondant à la position candidate et insérer un allèle de mutation causale sur chaque branche de cet arbre partiel ;
- calculer la statistique du χ^2 à l'aide du tableau de contingence ayant la forme du tableau 6.2 après l'ajout de la mutation sur chaque branche, en utilisant le phénotype de chaque séquence de l'échantillon, ainsi que le génotype du marqueur choisi ;
- garder la plus grande valeur de la statistique du χ^2 pour chaque ARG ;
- la moyenne de ces M valeurs maximales pour chaque marqueur est considérée comme sa statistique du χ^2.

La position du TIV est estimée par celle qui a la plus grande valeur de la statistique

Phénotype : ϕ \ Allèle : A	Mutant : 1	Non-mutant : 0	Totaux
Cas : 1	n_1	n_2	590
Témoin : 0	n_3	n_4	590
Totaux	$n_1 + n_3$	$n_2 + n_4$	1 180

Tableau 6.2: Tableau utilisé pour calculer la statistique χ^2 d'un marqueur de l'échantillon après l'ajout d'une mutation causale dans une branche d'un ARG. L'allèle associé au caractère d'intérêt est noté par «A». A=0 correspond à l'allèle non-muté et A=1 correspond à l'allèle muté.

du χ^2, T_{c_T} ; la formule de calcul de T_{c_T} est donc :

$$T_{c_T} = \frac{1}{M} \sum_{m=1}^{M} \max_{b \in B_{c_T}^m} T_{c_T}^b, \tag{6.1}$$

où $B_{c_T}^m$ est l'ensemble des branches de l'arbre partiel pour la position c_T, et b est une branche.

Pour chaque position candidate, on construit $M = 50$ ARGs pour calculer la statistique du χ^2. Il a été démontré que cette valeur est suffisante pour obtenir une bonne approximation de la fonction de vraisemblance (Descary, 2012), donc on la retient aussi pour calculer T_{c_T}.

6.3.2 Illustration des résultats obtenus avec la méthode DMap

Dans cette sous-section, nous illustrons les résultats obtenus avec la méthode DMap selon les différents scénarios donnés au tableau 6.1. Rappelons que seulement une partie de génotypes des L marqueurs est observée, et la méthode DMap peut s'effectuer avec des données incomplètes (pas besoin d'imputation). Sur chaque figure, on présente des graphiques représentant les résultats par rapport à la proportion de données manquantes. L'axe des x représente des

positions le long d'une séquence génétique de notre échantillon d'étude ; l'axe des y représente la statistique T_{c_T}. De plus, la ligne rouge verticale représente la position de TIV simulé ; le triangle vert montre la position estimée par la méthode χ^2 de DMap.

En observant les résultats obtenus, on constate que la méthode DMap donne de très bons résultats presque tout le temps. Cette excellente performance a lieu surtout quand le risque relatif est élevé ou quand la fréquence de la maladie est grande. Si ce n'est pas le cas, par exemple, au Scénario D_1, la statistique T_{c_T} calculée à différentes positions est plutôt homogène. Lorsque le risque relatif est très faible, la méthode DMap semble donner une estimation biaisée de la position du TIV (Scénario D_1 : 45% de données manquantes ; Scénario D_2 : 50% de données manquantes). On a réessayé cette méthode en augmentant le nombre de marqueurs à analyser jusqu'à 349, et on réussit à trouver la bonne position du TIV (voir Figures 6.15 et 6.16).

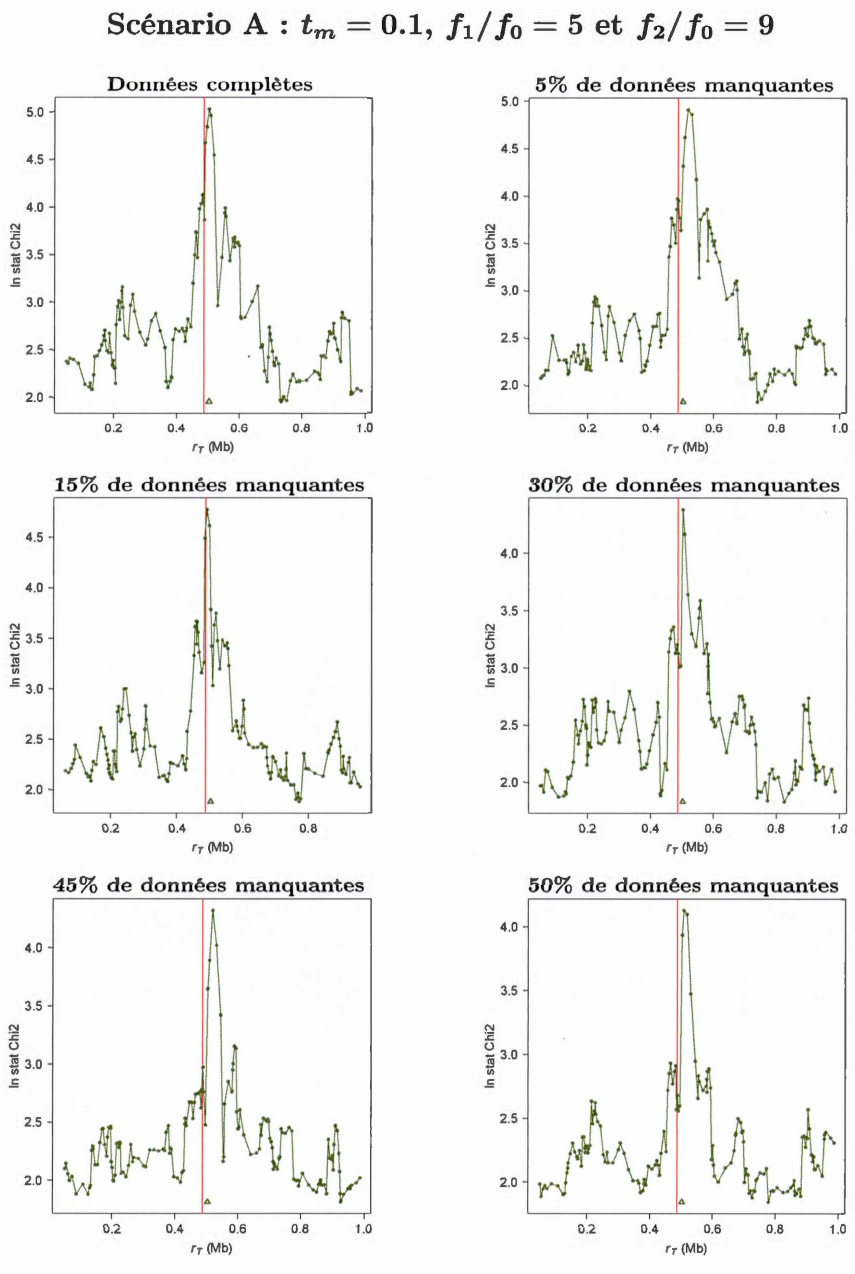


Figure 6.8: Illustration schématique des résultats de DMap avec 5 proportions différentes de données manquantes. L'échantillon est généré selon le scénario A.

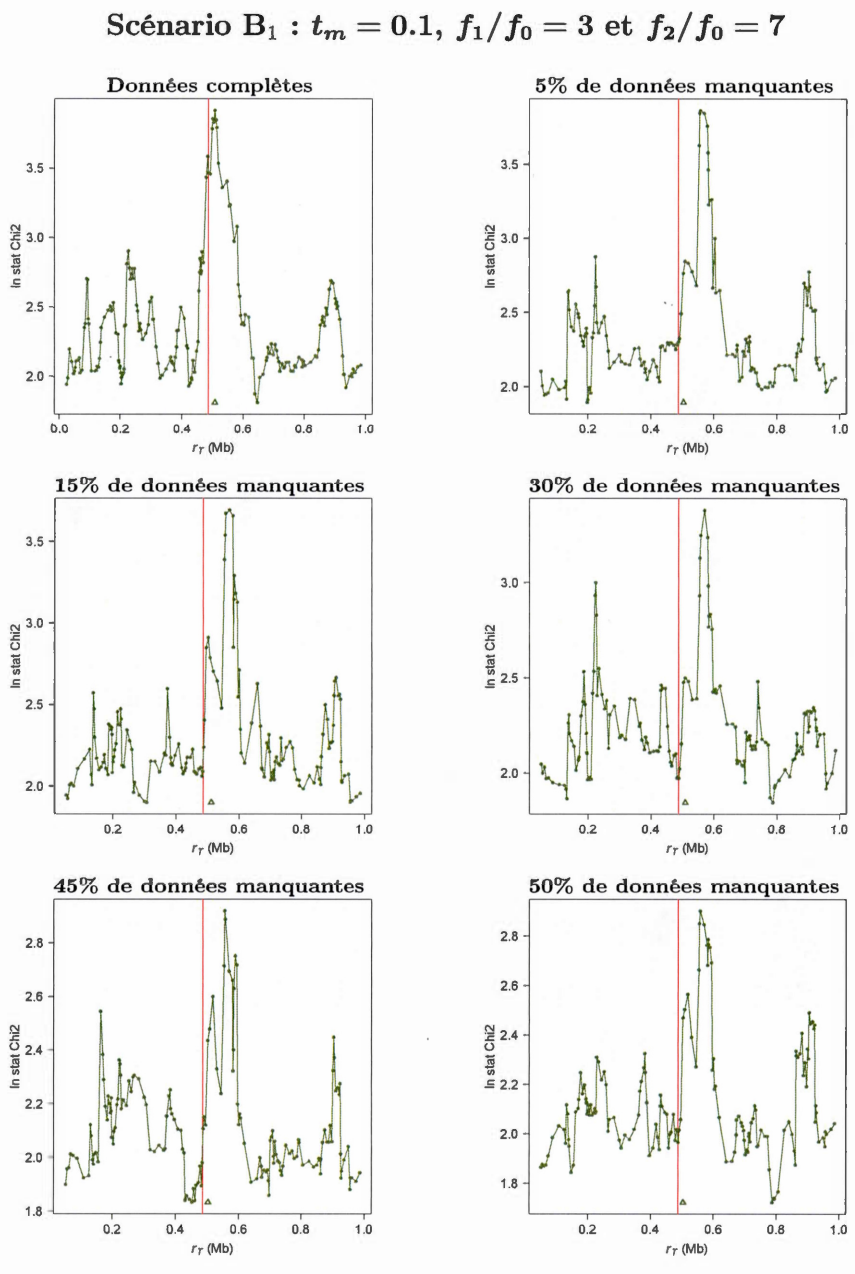


Figure 6.9: Illustration schématique des résultats de DMap avec 5 proportions différentes de données manquantes. L'échantillon est généré selon le scénario B_1.

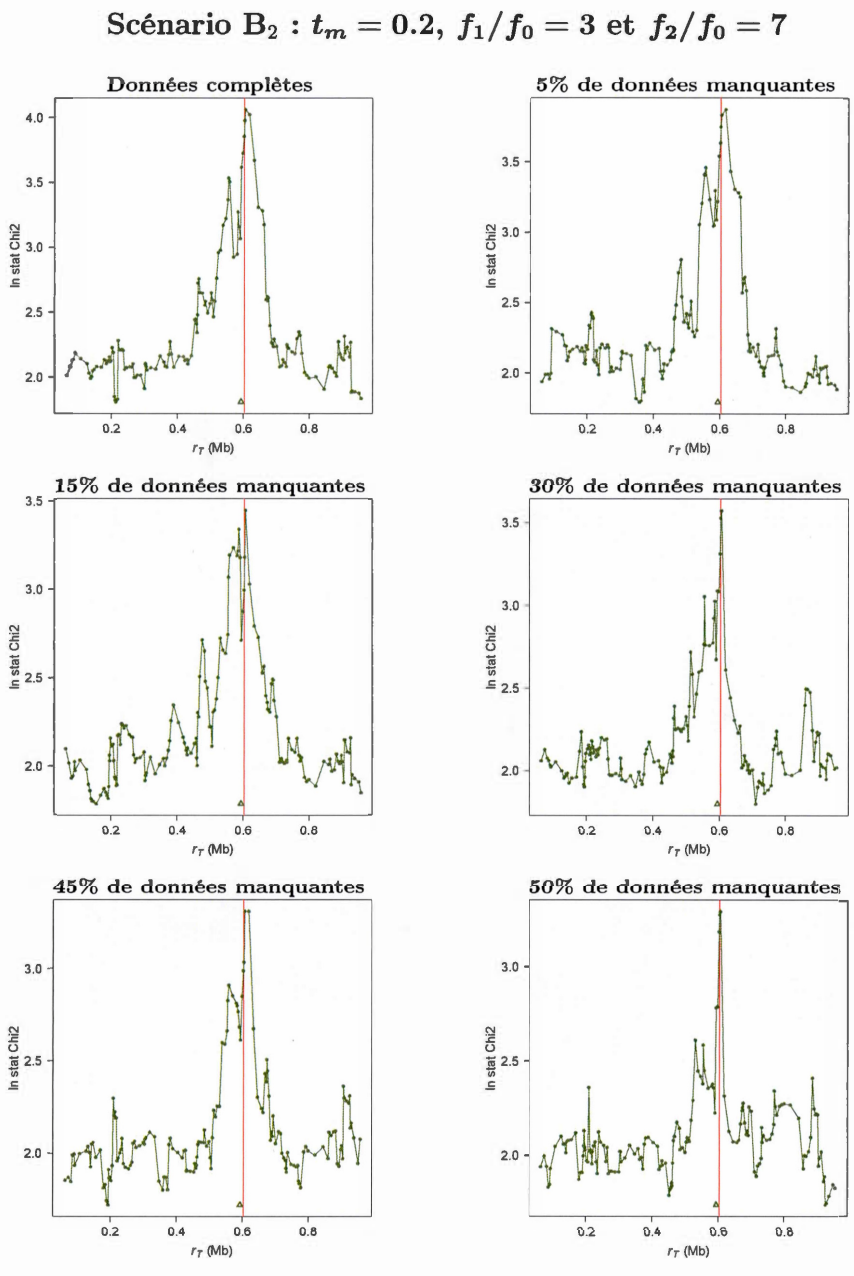


Figure 6.10: Illustration schématique des résultats de DMap avec 5 proportions différentes de données manquantes. L'échantillon est généré selon le scénario B_2.

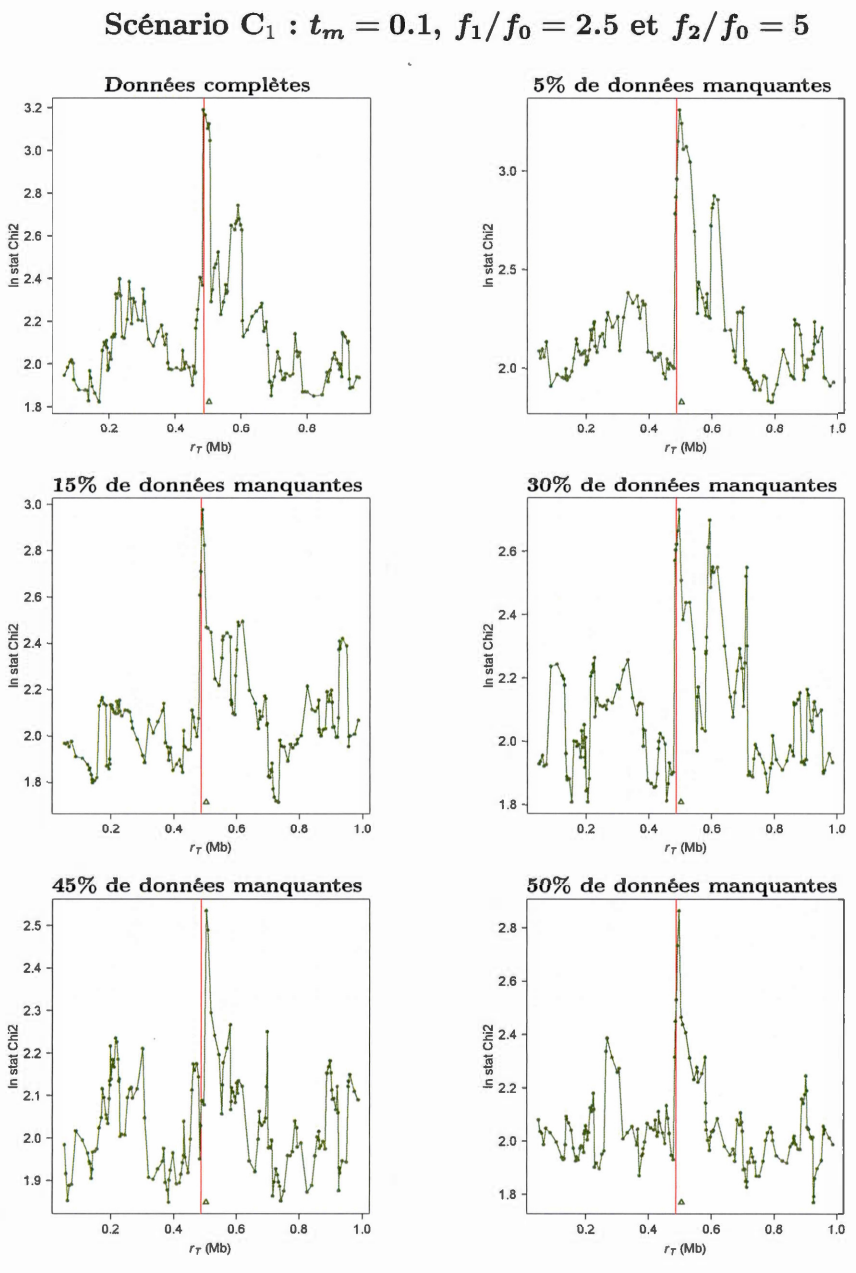


Figure 6.11: Illustration schématique des résultats de DMap avec 5 proportions différentes de données manquantes. L'échantillon est généré selon le scénario C_1.

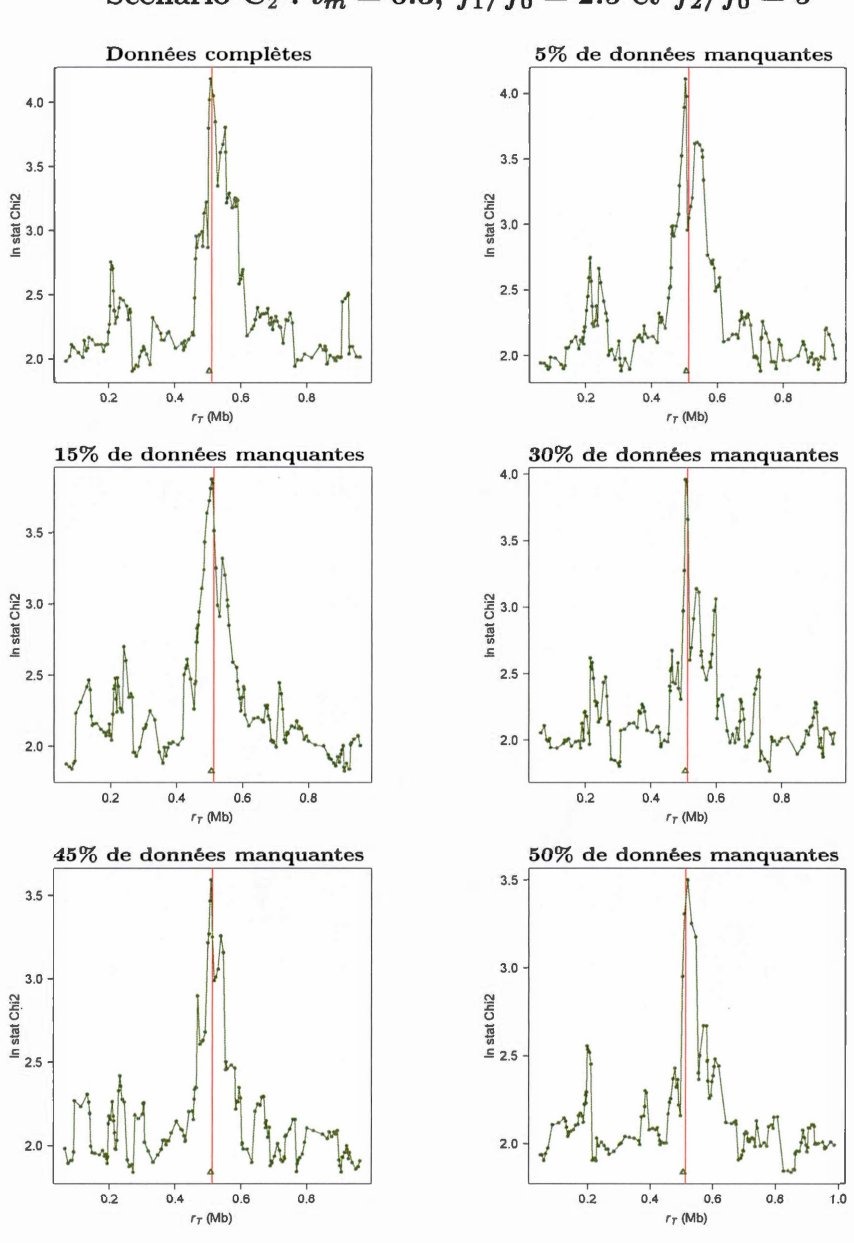


Figure 6.12: Illustration schématique des résultats de DMap avec 5 proportions différentes de données manquantes. L'échantillon est généré selon le scénario C_2.

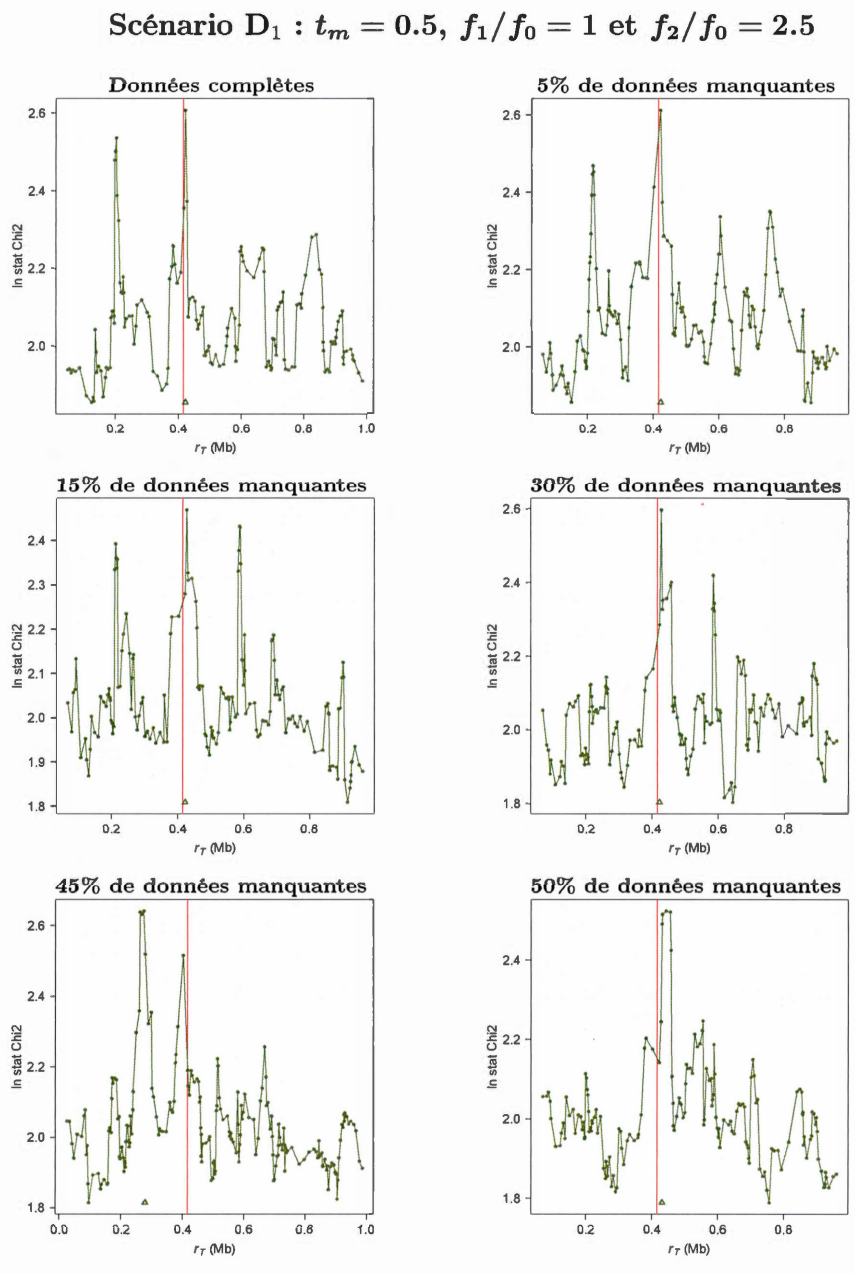


Figure 6.13: Illustration schématique des résultats de DMap avec 5 proportions différentes de données manquantes. L'échantillon est généré selon le scénario D_1.

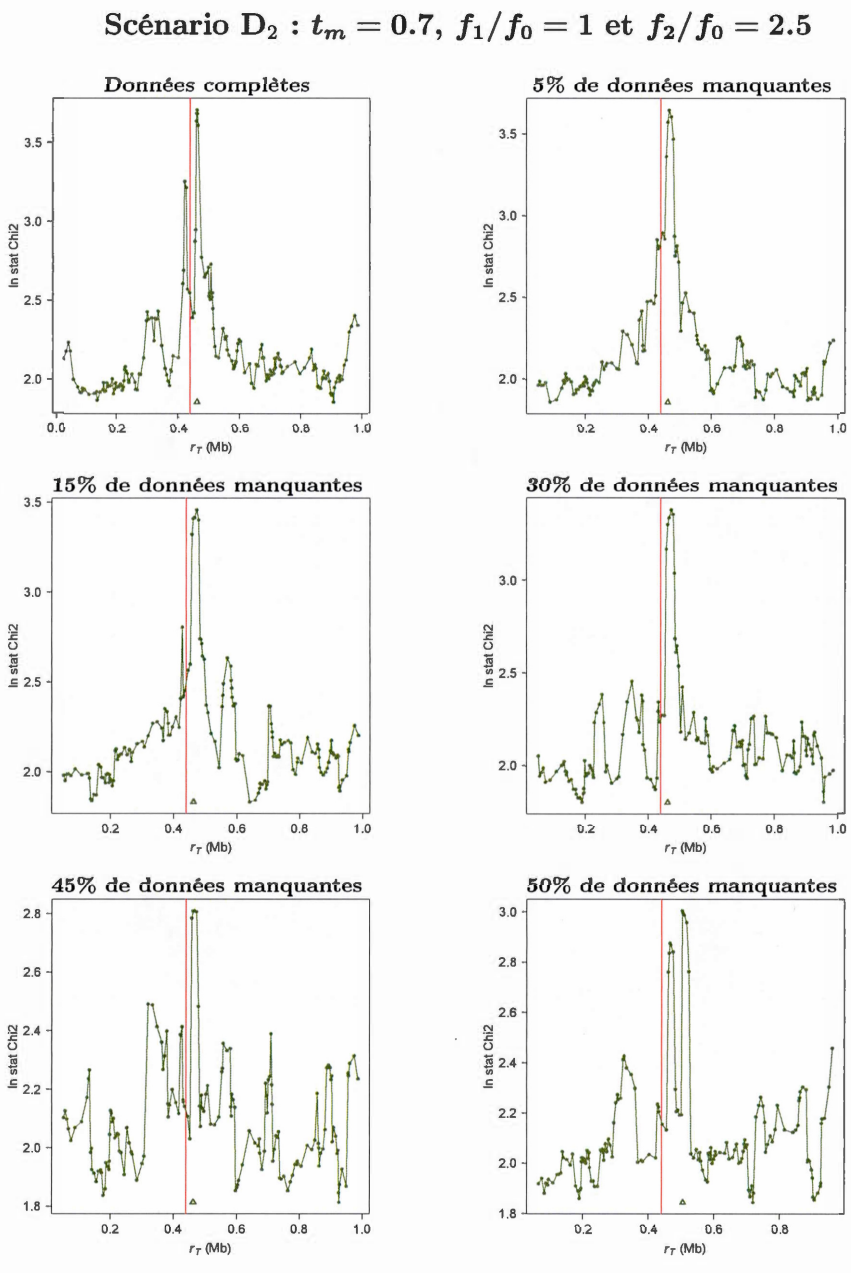


Figure 6.14: Illustration schématique des résultats de DMap avec 5 proportions différentes de données manquantes. L'échantillon est généré selon le scénario D_2.

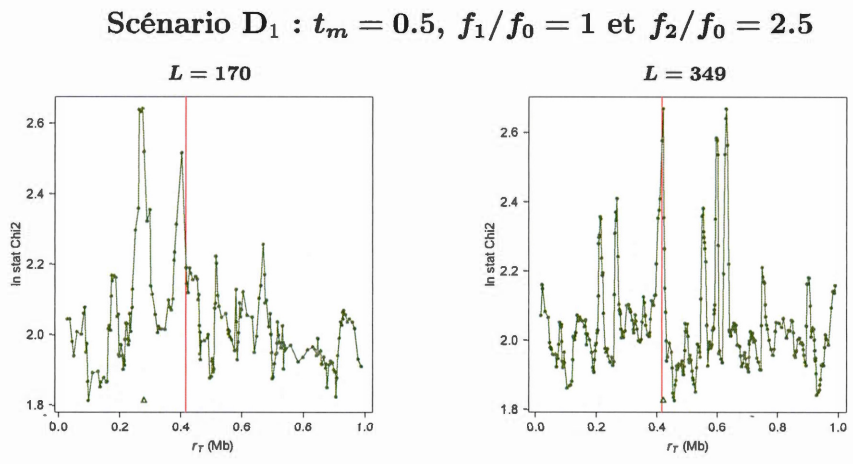


Figure 6.15: Illustration schématique des résultats de DMap avec 45% de données manquantes dans le scénario D_1. Les graphiques représentent les résultats avec 2 nombres différents de marqueurs ($L = 170$ et $L = 340$).

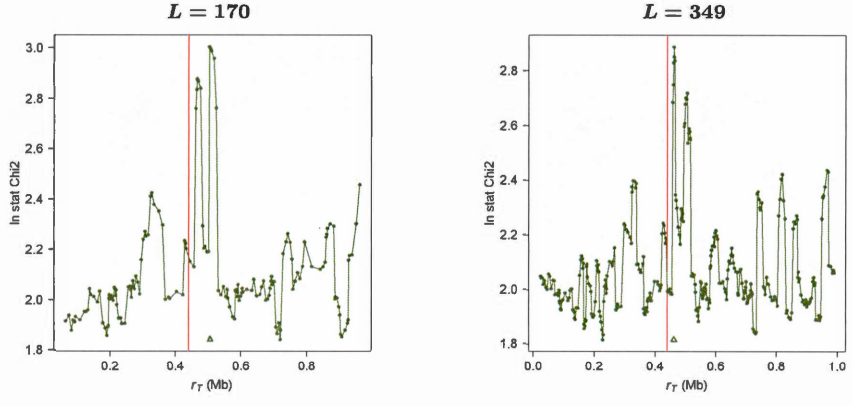


Figure 6.16: Illustration schématique des résultats de DMap avec 50% de données manquantes dans le scénario D_2. Les graphiques représentent les résultats obtenus avec 2 nombres différents de marqueurs ($L = 170$ et $L = 340$).

6.4 Discussion

En comparant les résultats obtenus par les deux méthodes, on remarque que la méthode DMap fonctionne aussi bien que la méthode SNPTEST. Dans certaines situations, la méthode DMap semble donner une meilleure estimation par rapport à la méthode SNPTEST. Par exemple, lorsque la proportion de données manquantes est importante (et presque pour tous les scénarios), la méthode DMap estime mieux la position du TIV que la méthode SNPTEST. En sachant que la méthode DMap utilise seulement une partie de données (entre 50% et 95%) pour effectuer les analyses, ces résultats sont remarquables. De plus, on croit que la performance de la méthode DMap pourrait être améliorée si on augmentait le nombre de marqueurs analysés et/ou le nombre d'ARG construits pour chaque marqueur. Il serait donc intéressant d'expérimenter plus d'options sur les paramètres utilisés dans la méthode DMap.

Un désavantage de la méthode DMap est qu'on doit connaître la fonction de pénétrance. Quant à la méthode SNPTEST, elle a besoin d'un panel de référence pour imputer les données manquantes. Grâce à l'agrandissement des projets HapMap et 1 000 Génomes, il y a de plus en plus d'informations génétiques sur différentes populations du monde qu'on peut utiliser pour l'imputation. L'estimation des génotypes manquants devient donc de plus en plus aisée, ce qui facilite l'utilisation de la méthode SNPTEST.

CONCLUSION

L'objectif de ce mémoire était de comparer l'efficacité de deux méthodes permettant de tester le lien entre un caractère d'intérêt et des marqueurs génétiques avec des données incomplètes. La première méthode, DMap, analyse directement les données incomplètes ; cette méthode suppose que les haplotypes sont connus et tient compte de la dépendance ancestrale entre les individus de l'échantillon. De ce fait, DMap peut se dispenser de l'imputation des données. La deuxième méthode, SNPTEST, effectue des analyses en faisant appel à l'imputation des données et prend en considération l'incertitude d'imputation.

Nous avons évalué et comparé la performance de ces deux méthodes par une étude de simulation. Sept scénarios ont été proposés pour générer différents échantillons, et ces scénarios varient par rapport au risque relatif et à la fréquence de la maladie. Pour chaque scénario, 5 proportions ont été choisies afin de simuler des données manquantes : on considère que 5%, 15%, 30%, 45% et 50% de génotypes sont manquants. D'après les résultats obtenus, les deux méthodes ont montré une sensibilité au risque relatif et à la fréquence de la maladie. Malgré cela, les deux méthodes semblent très efficaces quand il n'y a pas plus de 45% de données manquantes dans l'échantillon. Lorsqu'il y a une moitié de données manquantes dans l'échantillon, la méthode SNPTEST a de la difficulté à donner une bonne estimation de la position du TIV, tandis que la méthode DMap réussit à trouver la vraie position du TIV si l'on augmente le nombre de marqueurs à analyser. Ainsi, si la fonction de pénétrance est connue, il serait préférable d'utiliser la méthode DMap quand la proportion de données manquantes atteint 50%.

BIBLIOGRAPHIE

Cox, D. et Hinkley, D. (1979). *Theoretical Statistics*. Chapman and Hall/CRC.

Descary, M.-H. (2012). *DMAP : une nouvelle méthode de cartographie génétique fine adaptée à des modèles génétiques complexes*. (Mémoire de maîtrise). Université du Québec à Montréal.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. et Foll, M. (2013). Robust demographic inference from genomic and snp data. *PLOS Genetics*, *9*(10), 1–17.

Excoffier, L. et Foll, M. (2011). fastsimcoal : a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*(9), 1332–1334.

Fearnhead, P. et Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, *159*(3), 1299–1318.

Fisher, R. A. (1923). On the dominance ratio. *Royal Society of Edinburgh*, *42*, 321–341.

Garthwaite, P. H., Jolliffe, I. T. et Jones, B. (1995). *Statistical Inference*. Oxford University Press.

Hackett, C. et Broadfoot, L. (2003). Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity (Edinb)*, *90*(1), 33–38.

Hill, W. et Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and applied genetics*, *38*, 226–231.

Howey, R. et Cordell, H. J. (2012). PREMIM and EMIM : tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinformatics*, *13*, 149.

Howie, B. N., Donnelly, P. et Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics*, *5*(6), 1–15.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, *23*(2), 183–201.

Johnson, G. C., Esposito, L., Barratt, B. J. *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics*, *29*, 233–237.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, *61*(4), 893–903.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Processes and Applications*, *13*, 235–248.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of Applied Probability*, *19*, 27–43.

Kingman, J. F. C. (2000). Origins of the coalescent : 1974-1982. *Genetics*, *156*(4), 1461–1463.

Larribe, F. et Dupont, M. (2016). SimHGD : simulate case/control data for human genetic complex diseases. *Bioinformatics*.

Larribe, F., Lessard, S. et Schork, N. J. (2002). Gene mapping via the ancestral recombination graph. *Theoretical Population Biology*, *62*(2), 215–229.

Lewontin, R. C. et Kojima, K.-i. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, *14*(4), 458–472.

Li, N. et Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, *165*(4), 2213–2233.

Little, R. J. A. et Rubin, D. B. (2002). *Statistical Analysis with missing data* (2^e éd.). John Wiley & Sons, New Jersey.

Marchini, J. et Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511.

Marchini, J., Howie, B., Myers, S., McVean, G. et Donnelly, P. (2007). A new multipoint method for genome-wide association studies via imputation of genotypes : Supplementary methods. *39*(7), 906–913.

Minichiello, M. J. et Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Humain Genetics*, *79*, 910–922.

Reich, D. E. et Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in genetics*, *17*(9), 502–510.

The International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–58.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, *437*, 1299–1320.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, *449*(7164), 851–861.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, *16*(2), 97–159.