CARTOGRAPHIE GÉNÉTIQUE FINE : ÉVALUATION D'UNE MÉTHODE D'ESTIMATION DES ALLÈLES ET DU MODÈLE DE PÉNÉTRANCE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MATHIEU DUPONT

AVRIL 2013

Aux étudiants indignés de ce printemps québécois 2012, que nous n'oublierons jamais

ÉDITIONS ÉLECTRONIQUES

Le présent mémoire est optimisé pour être consulté en version électronique. Le texte contient des liens (en bleu) qui permettent d'accéder rapidement aux pages auxquelles il se réfère. De plus, 4 *boutons* simples, situés au bas des pages, permettent au lecteur de naviguer facilement et rapidement entre les diverses parties de l'ouvrage. Le bouton le plus utile est sans doute le deuxième ($<$), qui permet à tout moment de revenir à la page où l'on se trouvait précédemment. Le suivant ($>$), présent dans les chapitres, amène aux notations mathématiques, situées à la fin de chacun des chapitres. Enfin, les boutons $\ll$ et $\gg$ permettent d'accéder directement à la TABLE DES MATIÈRES et à l'INDEX, respectivement.

$\ll$ $<$ $\gg$

Deux versions électroniques sont disponibles : l'une classique (format lettre) et l'autre en format paysage, optimisée pour les tablettes numériques. Toutes deux peuvent être téléchargées à partir du site Web de mon directeur de recherche, Fabrice Larribe :

http://www.math.uqam.ca/fabricelarribe/memoires/MathieuDupont2012-FormatTabletteElectronique.pdf

http://www.math.uqam.ca/fabricelarribe/memoires/MathieuDupont2012-FormatLettre.pdf

Quoiqu'il s'agisse de fichiers en format PDF, les graphiques sont optimisés pour être lus avec le logiciel *Aperçu* (*Preview*).

REMERCIEMENTS

Je remercie Fabrice Larribe, mon directeur de recherche, pour son enthousiasme et sa passion, son ouverture et sa confiance, son dévouement et sa disponibilité, sa créativité inspirante et son sens de l'innovation. Ce fut toujours un très grand plaisir de travailler avec toi, et d'entretenir de longues discussions interminables !

Je remercie Sorana Froda et Karim Oualkacha, membres du jury, pour leurs corrections et commentaires pertinents. Merci Sorana pour votre confiance et vos encouragements constants.

Je remercie la communauté ConTEXt qui m'a été très utile pour implanter le modèle de présentation des mémoires et des thèses de l'UQAM en ConTEXt, incluant les liens de navigation (en bleu).

Je remercie mon cousin Jean-Philippe Boucher de m'avoir inspiré et encouragé à entreprendre ce projet, et pour sa vision sociale de l'université.

Finalement, je remercie Julie pour sa passion inspirante, sa confiance inébranlable et ses encouragements illimités.

TABLE DES MATIÈRES

LISTE DES FIGURES

Figure Page

LISTE DES TABLEAUX

RÉSUMÉ

Nous présentons et testons une méthode d'estimation des allèles d'une mutation potentiellement associée à un phénotype ainsi qu'une méthode d'estimation de son modèle de pénétrance. Ces deux méthodes reposent sur un algorithme EM et s'insèrent dans une méthode de cartographie génétique fine, *MapARG*, basée sur le processus de coalescence. La sensibilité des deux méthodes d'estimation aux risques relatifs du modèle de pénétrance réel de la mutation, à la taille des échantillons ainsi qu'à la largeur des fenêtres utilisées est systématiquement évaluée. Les deux méthodes s'avèrent performantes, particulièrement pour des risques relatifs forts. La taille des échantillons exerce peu d'influence, mais des fenêtres plus larges donnent de meilleurs résultats. L'estimation préalable du modèle de pénétrance montre un certain effet bénéfique sur l'estimation subséquente des allèles, comparativement à l'utilisation du vrai modèle connu. Aussi, la méthode d'estimation du modèle de pénétrance, basée sur une distance calculée entre les haplotypes primitifs et mutants, montre en soi un certain potentiel comme méthode de cartographie génétique.

MOTS-CLÉS : algorithme EM, cartographie génétique, coalescence, MapARG, modèle de pénétrance, risque relatif, SNP, statistique génétique, vraisemblance composite

INTRODUCTION

Bien qu'*Homo sapiens* pratique la sélection artificielle sur d'autres espèces depuis des millénaires, parfois consciemment et parfois inconsciemment, ce n'est que depuis tout récemment dans son histoire qu'il en comprend en partie les mécanismes, grâce notamment aux travaux de Darwin (1859) et de Mendel (1865). Aujourd'hui, les bases de données génétiques contiennent des quantités astronomiques de ces données et en reçoivent continuellement de nouvelles. Divers champs de recherche tentent d'extirper des informations pertinentes de toutes ces données, comme la cartographie génétique, qui inclut le développement et la perfection de méthodes servant à identifier des mutations causales de phénotypes.

Un obstacle encore non résolu dans la cartographie génétique est l'ignorance des allèles que portent les individus d'un échantillon, à la mutation recherchée. De plus, le problème est accentué si l'on ignore également le modèle de pénétrance de la mutation en cause. Cet ouvrage a pour objectif de décrire et évaluer une méthode d'estimation des allèles et

une méthode d'estimation du modèle de pénétrance, qui s'insèrent toutes deux dans une méthode de cartographie génétique fine, *MapARG*, développée par Fabrice Larribe et qui repose sur la statistique génétique.

Le chapitre I présente la génétique des populations, en introduisant quelques concepts de génétique et du processus de coalescence, pour finalement y situer la cartographie génétique. La méthode de cartographie génétique fine MapARG sera ensuite décrite en détail au chapitre II. Nous décrierons et évaluerons au chapitre III une méthode d'estimation des allèles d'une mutation, en supposant que nous connaissons *a priori* son modèle de pénétrance. Finalement, au chapitre IV, nous décrierons et évaluerons une méthode d'estimation du modèle de pénétrance, ainsi que son effet sur la méthode d'estimation des allèles.

CHAPITRE I

GÉNÉTIQUE DES POPULATIONS

La *génétique des populations* étudie la distribution des différents *allèles* de certains gènes dans des populations. Ces allèles sont sous l'influence de quatre principaux processus évolutifs : la sélection naturelle, la dérive génétique, les mutations et la migration. Fortement basé sur des concepts mathématiques et ayant des applications directes en épidémiologie, ce domaine de recherche permet de comprendre la transmission des maladies génétiques. Initiée dans les années 1920 à 1940 par le statisticien R.A. Fisher et les généticiens S. Wright et J. Haldane, la génétique des populations fait le pont entre la théorie de l'évolution, élaborée par Charles Darwin (1859), et les mécanismes génétiques, décrits par Gregor Mendel à la même époque (1865) mais longtemps restés incompris. En appliquant les principes fondamentaux de la génétique mendélienne à l'échelle des populations, cette discipline donne ainsi naissance au néo-darwinisme.

La *cartographie génétique*, qui sera abordée au chapitre II et dont traite ce mémoire, puise ses outils mathématiques dans la génétique des populations. Le présent chapitre décrit les

concepts et les termes de la génétique des populations qui seront nécessaires à la compréhension du développement de la méthode. Le lecteur peut se référer au lexique (page 209) afin d'obtenir rapidement la définition d'un terme précis. La section 1.1 introduit les concepts minimaux de génétique qui sont nécessaires à une bonne compréhension du problème. Le lecteur familier avec ces concepts peut sauter cette section. La section 1.2 introduit quand à elle la *théorie de la coalescence*, sur laquelle repose le méthode de cartographie génétique *MapARG*, qui est l'objet de ce mémoire. Les notations mathématiques du présent chapitre se trouvent à sa fin, page 36.

1.1 Génétique

Le moine et botaniste Johann Gregor Mendel est communément reconnu pour être le fondateur de la génétique. Vers 1860, il réalise un jardin expérimental dans la cour de son monastère et conçoit un plan d'expériences utilisant les pois et visant à expliquer les lois de

l'hybridation. Sa méthodologie rigoureuse et la précision de ses observations lui permettent de poser les bases théoriques de la génétique moderne (Mendel, 1865). Cependant, très peu de scientifiques de son temps comprennent alors la formalisation mathématique de ses expériences. La présente section décrit les bases minimales de la génétique moderne qui seront nécessaires à la compréhension du problème, du niveau moléculaire à celui d'une population d'humains.

1.1.1 ADN : base de données biochimique

L'acide désoxyribonucléique (*ADN*), dont la structure moléculaire est découverte en 1953 par R. Franklin, M. Wilkins, F. Crick et J. Watson, est une très longue molécule linéaire qui contient l'information génétique d'un organisme. Il est constitué d'une séquence des quatre types de nucléotides, soient l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T), comme le montre ce court extrait :

...GCTTATCCTTACGGATAGCGGGGGCCGATGTAAAATACACATAACTGGGCAGGT...

L'ADN humain est long de plus de 3 milliards de nucléotides. La séquence précise de ces nucléotides renferme l'information génétique, encodée dans l'ADN, et nécessaire à la synthèse des molécules fonctionnelles dont un organisme est constitué. L'ensemble de l'ADN d'une espèce ou d'un individu est appelé *génome*. À l'exception de quelques types de cellules, notamment les globules rouges et les cellules sexuelles, toutes les *cellules nucléées* d'un organisme contiennent une copie identique de tout son génome. Le génome humain contient approximativement 20 000 gènes qui occupent seulement environ 2 % de la séquence nucléotidique, une grande portion du reste de l'ADN étant probablement vestigial.

Un *gène* est une section précise de l'ADN, souvent longue de plusieurs milliers de nucléotides, et qui code pour la synthèse d'une protéine ou d'un acide ribonucléique (ARN) fonctionnel. Selon la séquence exacte de nucléotides portée par un gène, la molécule pour laquelle ce gène code peut différer, ainsi que sa fonction. La substitution d'un seul nucléotide dans une

région clé d'un gène peut affecter significativement la fonction de la molécule résultante. La version d'un gène (la séquence précise qu'il porte) est appelée *allèle*. On réfère aussi à l'allèle d'un *marqueur* (à un locus précis, discuté à la section 1.1.3), qui peut être aussi court qu'un seul nucléotide (au plus 4 allèles possibles dans ce cas). En théorie, le nombre d'allèles possibles pour un gène de longueur n nucléotides est donc égal à 4^n. Le *génotype* d'un individu est l'ensemble de ses allèles pour un ensemble précis de gènes, de marqueurs ou pour une séquence précise de son génome, alors que son *phénotype* est le résultat de l'expression de son génotype, influencé par l'environnement. Ainsi, le fait d'être atteint ou pas d'une maladie constitue un exemple de phénotype binaire, et l'allèle que l'on possède pour un gène responsable de cette maladie, un exemple de génotype.

1.1.2 Recombinaison génétique : brassage aléatoire

Dans son incontournable ouvrage *The selfish gene* (1976), Richard Dawkins décrit un or-

ganisme vivant comme une machine complexe fabriquée par son génome dans le seul but de maximiser sa propre multiplication (du génome). Une des plus anciennes méthodes de multiplication est la simple *division cellulaire*, encore utilisée aujourd'hui par les organismes unicellulaires comme les bactéries ainsi que par les cellules des organismes pluricellulaires, et qui consiste à se dupliquer. La *reproduction sexuée*, apparue il y a plus d'un milliard d'années, permet cependant aux organismes qui la pratiquent de s'adapter plus rapidement à leur environnement changeant, en générant un *brassage des allèles*. Ainsi, de nouvelles combinaisons de gènes sont créées, permettant l'apparition de nouveaux caractères qui seront sélectionnés. La reproduction sexuée crée un brassage aléatoire mais organisé du génome, sur lequel repose toute la base de la génétique des populations. Ce brassage est effectué par deux processus décrits plus bas : (1) les *recombinaisons inter-chromosomiques* et (2) les *recombinaisons intra-chromosomiques*. Certains concepts de la reproduction sexuée doivent tout d'abord être compris afin de bien saisir l'importance du brassage génétique.

L'ADN d'une cellule humaine, déroulé, fait plus de 1 mètre de long. Cependant, lors de la division cellulaire, il est enroulé sur lui-même de façon très ordonnée, sous forme de *chromosomes*, longs d'environ 10 μm. À l'exception du court segment d'ADN mitochondrial, ne contenant que 37 gènes, toujours transmis par la mère et qui ne sera pas discuté dans cet ouvrage, le génome humain est divisé en 23 chromosomes retrouvés dans toutes les cellules nucléées. Chacun des quelques 20 000 gènes humains est situé à un locus précis sur l'un de ces 23 chromosomes.

La *ploïdie* d'une cellule (ou d'un organisme) fait référence au nombre d'exemplaires qu'elle contient de chacun de ces chromosomes. Les humains étant diploïdes, leurs cellules nucléées contiennent deux exemplaires de chacun des 23 chromosomes humains : l'un hérité de la mère et l'autre du père. Les deux chromosomes d'une même paire sont dits *homologues* ; ils contiennent les mêmes gènes, mais pas nécessairement les mêmes allèles.

La *méiose* est le processus biologique au cours duquel sont produits les gamètes (sperma-

tozoïdes ou ovules), cellules *haploïdes*, ne contenant que 23 chromosomes, soit un seul exemplaire de chaque chromosome. La fusion subséquente d'un spermatozoïde (haploïde) avec un ovule (haploïde) créera la première cellule (diploïde) d'un nouvel individu. Cette nouvelle cellule diploïde se multipliera ensuite par division cellulaire jusqu'à former un nouvel individu.

(1) Recombinaisons *inter*-chromosomiques

Au cours de la méiose, une cellule diploïde se divise en deux cellules haploïdes, chacune d'elles ne conservant que 23 chromosomes, soit un seul des deux chromosomes de chacune des 23 paires de chromosomes homologues, aléatoirement et indépendamment des 22 autres paires de chromosomes homologues. Ainsi, à elle seule, cette ségrégation indépendante des chromosomes lors de la méiose peut générer, à partir du bagage génétique d'un individu, c'est-à-dire à partir des chromosomes homologues qu'il a lui-même hérités de ses parents,

pas moins de 2^{23} combinaisons différentes !

(2) Recombinaisons *intra*-chromosomiques

Afin que chaque gamète obtienne exactement l'un des deux chromosomes de chacune des 23 paires de chromosomes homologues, ceux-ci sont physiquement regroupés par paires avant la division cellulaire. À ce moment, alors que les deux chromosomes homologues d'une paire sont alignés et très rapprochés l'un de l'autre, il peut se produire un phénomène d'*enjambement* (*crossover*). Lors de ce phénomène, les deux chromosomes s'attachent aléatoirement en un point précis le long de la séquence (le même locus sur les deux chromosomes), puis se détachent. Cependant, la partie *gauche* (étiquette arbitraire) du chromosome M (hérité de la mère) se retrouve alors avec la partie *droite* du chromosome P (hérité du père), plutôt qu'avec la partie *droite* du chromosome M. Les deux chromosomes se trouvent alors à avoir été croisés (*crossover*) en un point aléatoire. Les gamètes résultants ne possèderont alors

pas aléatoirement soit le chromosome hérité de la mère ou bien celui provenant du père, mais plutôt une *re*-combinaison aléatoire des deux.

Il est à noter que plus d'un enjambement peuvent se produire entre deux chromosomes homologues, résultant en deux chromosomes plusieurs fois recombinés (en plusieurs points de recombinaison). Ainsi, si un nombre pair d'enjambements se produisent entre deux gènes, les allèles de ces deux gènes se trouveront alors conservés ensemble sur les chromosomes résultants.

Les mécanismes de brassage génétique reliés à la reproduction sexuée décrits ici, soient les recombinaisons *inter*- et *intra*- chromosomiques, impliquent trois générations, soient (1) les parents du reproducteur, (2) le reproducteur et (3) les enfants du reproducteur. Toutes les cellules du reproducteur contiennent les mêmes 46 chromosomes, soit, pour chacune des 23 paires, un chromosome hérité de sa mère, et l'autre de son père. Chaque gamète qu'il produira contiendra 23 chromosomes, c'est-à-dire, aléatoirement pour chacune des

23 paires, soit le chromosome qu'il a lui-même (le reproducteur) hérité de sa mère, soit celui provenant de son père, ou bien encore une *re*-combinaison des deux (recombinaison *intra*-chromosomique).

1.1.3 SNP : unité statistique

Le génome humain ne montre aucune variation, ou *polymorphisme* (coexistence dans une population de plusieurs allèles pour un gène ou un nucléotide), sur plus de 99 % de sa séquence nucléotidique. Autrement dit, tous les individus possèdent exactement la même séquence sur plus de 99 % du génome. La variation qui existe est causée par des mutations qui surviennent lors de la méiose et qui permettent au génome d'évoluer et de s'adapter à son environnement changeant. Ces mutations comportant un important aspect aléatoire, il n'est pas rare qu'elles aient un impact néfaste sur le phénotype d'un organisme, contribuant parfois à causer certaines maladies.

De plus en plus de ces polymorphismes sont maintenant bien documentés et constituent des *marqueurs génétiques* dont on connait les loci précis dans le génome ainsi que leurs différents allèles retrouvés dans la population. Les *SNPs* (*single-nucleotide polymorphisms*), prononcés « snip », forment la majorité de l'ensemble des variations génétiques humaines et constituent la forme la plus simple de marqueurs génétiques. Ils sont constitués d'un seul nucléotide et presque tous ont seulement deux allèles possibles (sur une possibilité théorique de $4^1 = 4$). Deux SNPs sur trois ont les allèles C et T. Il est important de savoir qu'une grande partie des SNPs n'ont aucune influence sur le phénotype de l'individu porteur, soit parce qu'ils se trouvent dans une région *intergénique* de l'ADN (pas dans un gène), ou bien pour des raisons biochimiques qu'il est inutile de décrire ici.

La figure 1.1 montre un exemple de SNPs sur une courte séquence tronquée du chromosome Y de 5 individus. Chez l'humain, le chromosome 23 est aussi appelé chromosome sexuel, car il détermine le sexe de l'individu. Les femmes possèdent deux chromosomes homologues

Figure 1.1 Polymorphisme nucléotidique. Séquence homologue située sur le chromosome Y, chez 5 individus. Trois SNPs consécutifs sont mis en évidence et numérotés.

X, alors que les hommes n'en possèdent qu'un, associé avec un chromosome Y. Ainsi, un ovule contient toujours un chromosome X, et c'est le spermatozoïde qui détermine le sexe du nouvel individu formé, selon qu'il contient un chromosome X ou Y. Puisque les chromosomes X et Y ne recombinent pas sur 95 % de leur séquence et que les porteurs

d'un chromosome Y n'en possèdent qu'une seule copie, cette séquence d'ADN constitue un exemple simplifié pour comprendre certains concepts de base, faisant abstraction de la diploïdie du reste du génome, ainsi que des recombinaisons intra-chromosomiques.

1.2 Théorie de la coalescence

La méthodologie utilisée dans cet ouvrage (*MapARG*, chapitre II) se base sur des principes mathématiques de la *théorie de la coalescence*, initialement développée au début des années 1980 par John Kingman (1982, 2000). Le processus de coalescence est un processus stochastique qui permet de modéliser la généalogie d'individus dont on ne connait pas les liens de parenté. Les principes de base de la théorie seront exposés dans cette section, ainsi que ceux qui sont requis pour la compréhension du problème. Le lecteur qui désirerait approfondir ses connaissances sur le sujet peut se référer aux ouvrages suivants : Wakeley, 2009 et Hein *et al.*, 2005.

En génétique, la théorie de la coalescence est un modèle rétrospectif de la génétique des populations. Ce modèle retrace tous les allèles d'un gène partagé par tous les membres d'une population à un seul exemplaire ancestral, l'ancêtre commun le plus récent, ou *MRCA* (*Most Recent Common Ancestor*). Les relations d'hérédité entre les allèles sont généralement représentées par une généalogie, ressemblant à un arbre phylogénétique. La compréhension des propriétés statistiques de cette généalogie sous différentes hypothèses forme la base de la théorie de coalescence.

Des modèles de dérive génétique sont créés en reculant dans le temps. Dans le cas le plus simple, la théorie de la coalescence suppose l'absence de recombinaison, de sélection naturelle, de migration des gènes et de structure de la population. De récentes avancées permettent cependant d'étendre la théorie de base en y incluant de la recombinaison, de la sélection, et pratiquement n'importe quel modèle évolutif ou démographique dans l'analyse génétique de la population.

1.2.1 Nature des données

Les données prennent la forme de séquences homologues d'ADN, donc provenant de la même région génomique, comme par exemple un gène précis ou bien une section plus large mais précise d'un chromosome particulier. À partir de ces séquences génétiques observées aujourd'hui, nous cherchons à comprendre différents aspects de la généalogie qui les relie entre elles jusqu'à leur ancêtre commun.

La figure 1.2 montre à titre d'exemple une infime partie de la séquence d'ADN du chromosome Y (qui compte environ 50 millions de nucléotides) de 5 individus. Les séquences sont alignées, c'est-à-dire que les nucléotides homologues sont alignés les uns vis-à-vis des autres. Puisque seule la variation dans les séquences contient de l'information et est susceptible de nous intéresser, seuls les allèles des marqueurs, dans ce cas-ci des SNPs, sont conservés. Comme nous le verrons plus loin, la position précise de ces SNPs est également très importante dans la méthode. Aussi, puisque les SNPs sont binaires, les données peuvent

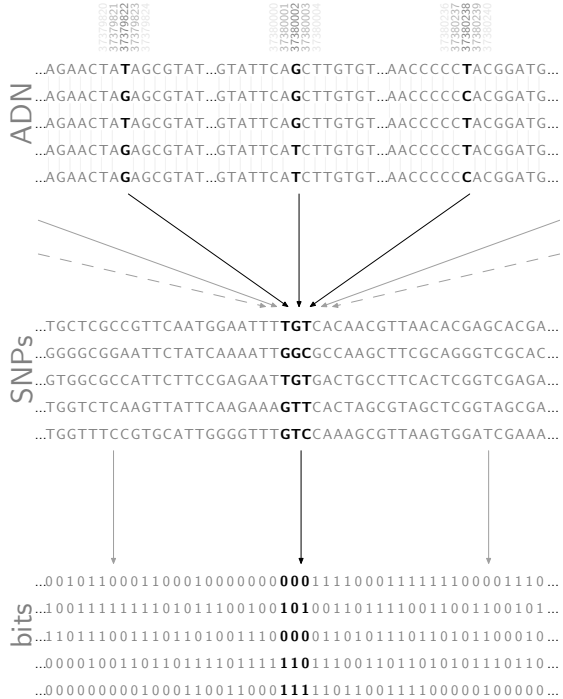


Figure 1.2 Structure et provenance des données. *Haut :* séquences homologues du chromosome Y de 5 individus. Les nucléotides autour des SNPs sont numérotés. *Milieu :* Seuls les SNPs, qui contiennent de la variation, sont conservés pour analyse. *Bas :* Les SNPs étant binaires, leurs allèles sont transformés en bits.

alors être transformées en bits. Comme nous le verrons plus loin, sous un *modèle de sites infinis* et sous l'hypothèse que nous connaissons l'allèle primitif, celui-ci prend la valeur 0 alors que l'allèle mutant prend la valeur 1.

Lorsque nous comparons des «séquences», il est sous-entendu qu'il s'agit de séquences homologues. Sauf lorsque précisé, il peut tout aussi bien s'agir de simples nucléotides, de courtes séquences (quelques nucléotides), de gènes ou même de chromosomes entiers. Il ne s'agit pas nécessairement de séquences de nucléotides consécutifs sur l'ADN, mais les séquences comparées sont toujours homologues, c'est-à-dire que chaque séquence contient les mêmes nucléotides (mêmes positions sur l'ADN).

1.2.2 Modèle de Wright-Fisher

R.A. Fisher (1930) et S. Wright (1931) ont développé un modèle simple pour décrire la relation généalogique qui relie plusieurs séquences homologues entre elles. Ce modèle est basé sur six hypothèses :

1. générations discrètes ;
2. individus haploïdes ;
3. taille constante de la population ;
4. absence de sélection (valeur adaptative égale des individus) ;
5. absence de structure géographique ou sociale de la population ;
6. absence de recombinaison.

Alors que, comme nous le verrons, l'hypothèse 6 est déjà assouplie dans le modèle utilisé par la méthode MapARG, l'objectif de cet ouvrage ne pourrait être atteint sans également

assouplir l'hypothèse 2, dont les difficultés seront étudiées au chapitre III. La figure 1.3 montre un exemple d'une généalogie qui suit le modèle de Wright-Fisher dans sa forme la plus simple, soit une population de taille *constante* de 12 séquences *haploïdes* qui meurent à la naissance de chaque nouvelle génération *discrète*. À partir des séquences observées aujourd'hui (génération du bas), on remonte dans le temps (vers le haut) en choisissant aléatoirement pour chaque séquence son parent parmi les séquences de la génération précédente (en haut de sa génération), avec remise.

1.2.3 Graphe de recombinaison ancestral

Le graphe de recombinaison ancestral, ou *ARG* (*Ancestral Recombination Graph*, Griffiths et Marjoram, 1996), outil fondamental de la méthode MapARG, est la construction mathématique d'une généalogie possible reliant des séquences observées à leur MRCA, en utilisant trois types d'évènements possibles. Ces trois évènements, soient la *coalescence*, la *mutation*

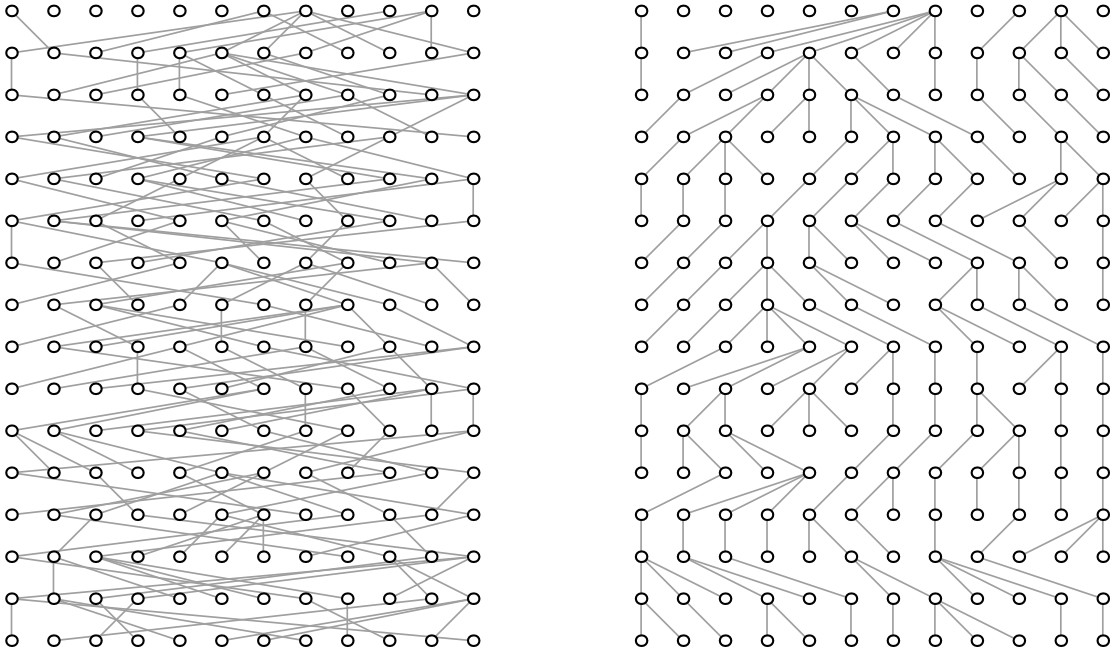


Figure 1.3 Exemple d'une généalogie suivant le modèle de Wright-Fisher. Les lignées sont démêlées à droite. Une population constante de 12 séquences haploïdes évolue sur 16 générations discrètes. Les points représentent les séquences et une rangée de points correspond à une génération. La généalogie se construit en remontant dans le temps (vers le haut) en associant à chaque séquence un parent parmi les séquences de la génération précédente (en haut), aléatoirement avec remise (segments gris).

et la *recombinaison*, seront ajoutés un à un à la description de notre modèle d'ARG.

1.2.3.1 Coalescence

Lorsque deux séquences ou plus choisissent le même parent, il y a coalescence de leurs *lignées*. La figure 1.4 montre une généalogie simulée des séquences de la génération observée, ainsi que son arbre de coalescence, dans lequel la longueur des branches correspond au temps d'occurrence des coalescences de deux lignées. Le nombre de lignées diminue à mesure qu'elles coalescent, jusqu'à ce que l'on atteigne le MRCA (en noir).

À chaque nouvelle génération (en remontant dans le temps), deux lignées données vont coalescer avec probabilité $\frac{1}{N}$, où N est la taille (constante, ici 12) de la population. Le temps avant qu'elles coalescent, en nombre de générations, suit donc une loi géométrique de paramètre $\frac{1}{N}$. En supposant N très grand ($\geq 10\,000$) et que seulement deux lignées peuvent coalescer à la fois, il existe $\binom{n}{2}$ couples possibles de lignées qui peuvent coalescer

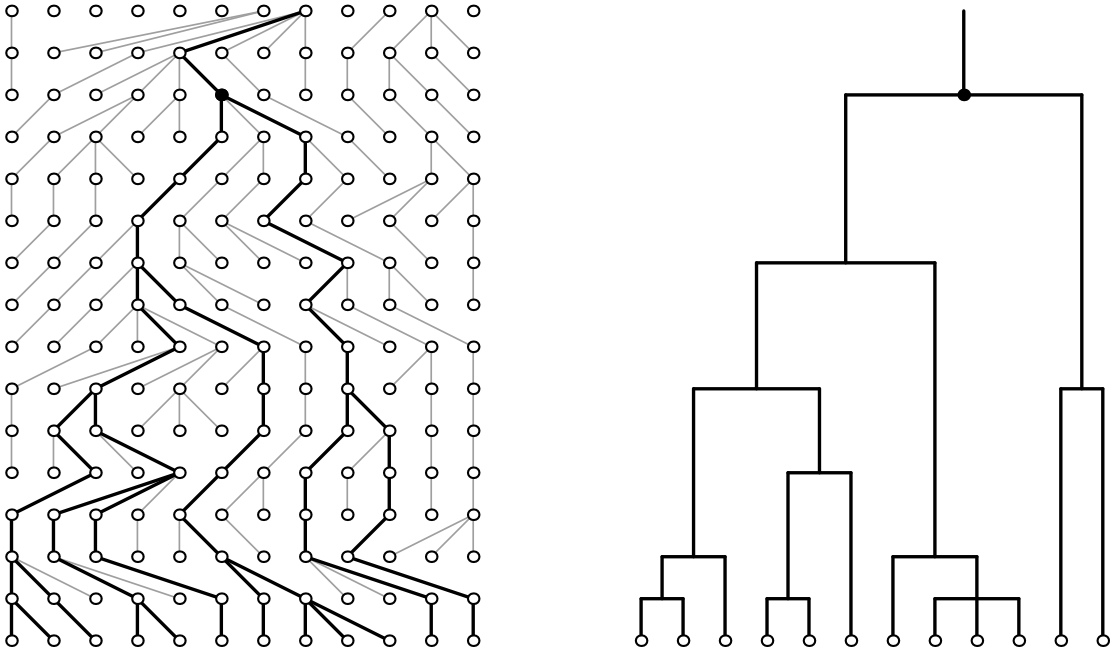


Figure 1.4 Arbre de coalescence. *Gauche :* La généalogie et le MRCA des séquences de la génération observée (la plus récente, en bas) sont en noir. *Droite :* Arbre de coalescence. La longueur des branches correspond au temps d'occurrence des coalescences de deux lignées.

lorsqu'il en reste n à relier. En notant alors T_C^n le temps (en nombre de générations) avant qu'il y ait coalescence de deux lignées, la probabilité que la prochaine coalescence se produise au temps k peut être approximée par

$$P(T_C^n = k) \;=\; \left[1 - \binom{n}{2}\frac{1}{N}\right]^{k-1} \binom{n}{2}\frac{1}{N} \,.$$

En génétique, un échantillon de n_e séquences provient habituellement d'une population de taille N beaucoup plus grande. En supposant $n \leq n_e \ll N$ et une échelle de temps continu $x = \frac{k}{N}$, cette distribution peut être approximée par une loi exponentielle de taux $\binom{n}{2}$:

$$P(T_C^n \leq x) \;=\; 1 - e^{-\binom{n}{2}x} \,.$$

Ainsi, pour simuler une généalogie reliant n_e séquences tirées d'une population de taille N à leur MRCA, il suffit, pour chacun des $n_e - 1$ évènements de coalescence nécessaires, de simuler un temps d'attente et de choisir aléatoirement une paire de lignées parmi les n restantes, jusqu'à ce qu'il n'en reste qu'une seule, correspondant au MRCA.

1.2.3.2 Mutation

Les mutations seront modélisées selon le modèle de mutations à sites infinis : nous assumons que les mutations sont des évènements rares et qu'elles ne peuvent se produire qu'une seule fois sur un SNP donné. Ainsi, en notant tous les SNPs du MRCA par 0, chacun des SNPs des séquences observées sera soit identique au SNP correspondant du MRCA et conservera l'allèle 0, soit il aura subi une mutation quelque part le long de sa lignée et prendra alors l'allèle 1 ; nous y reviendrons. De plus, rappelons que selon l'hypothèse 4 du modèle de Wright-Fisher, les mutations sont neutres et n'exercent donc aucune influence sur la structure de la généalogie.

Lorsqu'une mutation survient, *tous* les descendants et *seulement* les descendants de la séquence mutée porteront l'allèle 1 au SNP muté. Autrement dit, en construisant une généalogie en remontant dans le temps à partir d'un échantillon de séquences observées, pour que la mutation d'un SNP puisse être créée, il ne doit pas rester plus d'une seule lignée

portant l'allèle 1 à ce SNP. De la même manière, seuls les couples de lignées identiques, donc portant exactement les mêmes mutations, peuvent coalescer. La figure 1.5 illustre deux mutations dans la généalogie vue précédemment.

Les mutations surviennent de façon indépendante pour chacune des lignées. Ainsi, nous noterons μ la probabilité qu'une mutation se produise sur une lignée entre deux générations. Tout comme pour la coalescence, le temps, en nombre k de générations, avant qu'une mutation survienne lorsqu'il reste n lignées, noté T_M^n, suit une loi géométrique de paramètre $n\mu$:

$$P(T_M^n = k) \;=\; (1 - n\mu)^{k-1} \;\; n\mu \;,$$

qui peut aussi être approximée sur une échelle de temps continu $x = \frac{k}{N}$ par une loi exponentielle de taux $\frac{n\,\theta}{2}$:

$$P(T_M^n \leq x) \;=\; 1 - e^{-\frac{n\,\theta}{2}x} \;,$$

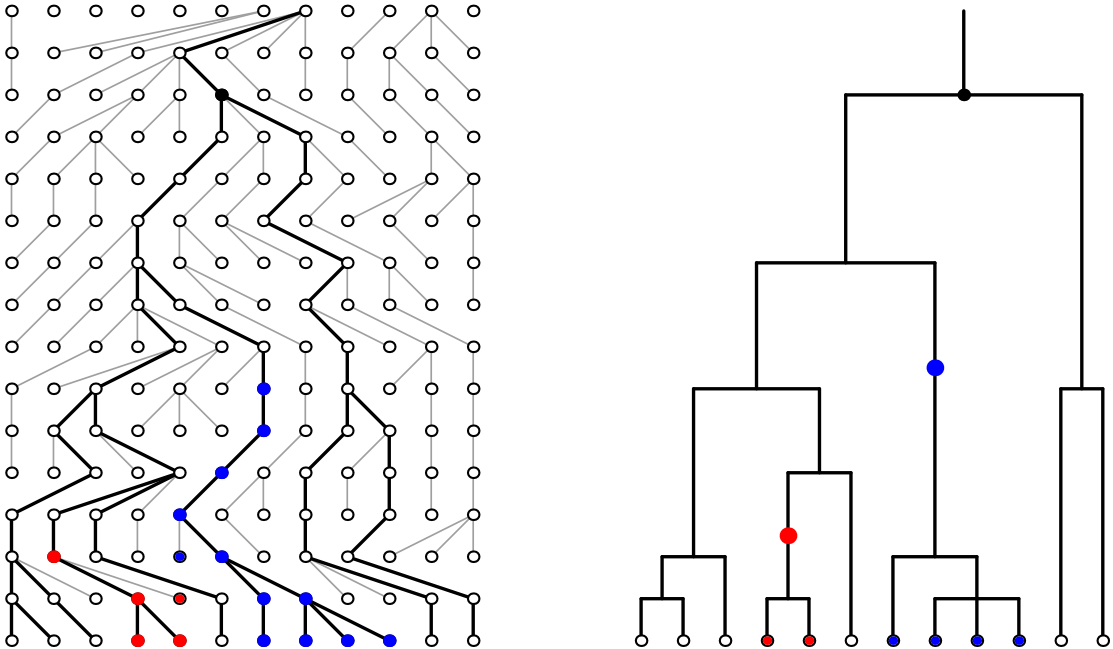


Figure 1.5 Arbre de coalescence avec mutations. *Gauche :* Deux mutations sont simulées dans la généalogie. Toutes les séquences d'une lignée qui sont situées en dessous de l'apparition d'une mutation portent cette dernière. *Droite :* Arbre de coalescence avec mutations. La hauteur des points de mutation correspond à leur temps d'occurrence.

où $\theta = 2\,\mu N$ est le taux de mutation dans la population.

Comme les coalescences et les mutations se produisent de façon indépendante, le temps T^n avant qu'un de ces évènements survienne suivra une loi exponentielle de taux

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

1.2.3.3 Recombinaison

La création d'un évènement de recombinaison dans un processus de coalescence est moins triviale qu'une coalescence ou une mutation. La recombinaison est cependant nécessaire à l'élaboration d'un modèle mathématique de généalogies compatible avec la réalité diploïde de la génétique humaine. Hudson propose dès 1983 l'inclusion des recombinaisons dans le processus de coalescence. Pour créer un tel évènement à partir d'une séquence (en remontant dans le temps), un site de recombinaison est choisi aléatoirement le long de la séquence.

Ainsi, la section de la séquence qui est située d'un côté du site de recombinaison proviendra d'un parent, et l'autre section, de l'autre parent. Chacune de ces deux nouvelles séquences partielles (les parents de la séquence) se verra complétée par une séquence temporaire partielle *non ancestrale*, c'est-à-dire ne provenant pas de l'éventuel MRCA. En fait, lorsqu'il n'y a pas de recombinaison entre une séquence et ses parents, c'est que toute la séquence provient d'un seul de ses parents, et on ne s'intéresse pas à l'autre parent, qui est non ancestral.

En notant r la probabilité qu'une recombinaison se produise sur une lignée entre deux générations, le temps, en nombre k de générations, avant qu'une recombinaison survienne lorsqu'il reste n lignées, noté T_R^n, suit une loi géométrique de paramètre nr :

$$P(T_R^n = k) \; = \; (1 - nr)^{k-1} \; nr \; ,$$

qui peut être, encore une fois, approximée sur une échelle de temps continu $x = \frac{k}{N}$ par une loi exponentielle de taux $\frac{n\,\rho}{2}$:

$$P(T_R^n \leq x) \;=\; 1 - e^{-\frac{n\,\rho}{2}x}\,,$$

où $\rho = 2\,r\,N$ est le taux de recombinaison dans la population. En incluant les recombinaisons, le temps T^n avant qu'un évènement surgisse suivra une loi exponentielle de taux

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

La figure 1.6 illustre un exemple d'ARG pour relier 4 séquences de 8 SNPs à leur MRCA. Dans cet ARG, 7 évènements sont nécessaires pour atteindre le MRCA. Une application de l'ARG consiste à en générer plusieurs qui sont compatibles avec un échantillon de séquences observées, et d'évaluer la vraisemblance de cet échantillon en conditionnant sur tous les ARGs possibles. Il est important de noter que, contrairement à l'arbre de coalescence, qui constitue un processus de mort, l'ARG est un processus de naissance et de mort. Conséquemment, il n'existe pas un nombre fini d'ARGs possibles pour relier un échantillon de séquences à leur MRCA. Étant donnée cette limitation, nous verrons au chapitre II com-

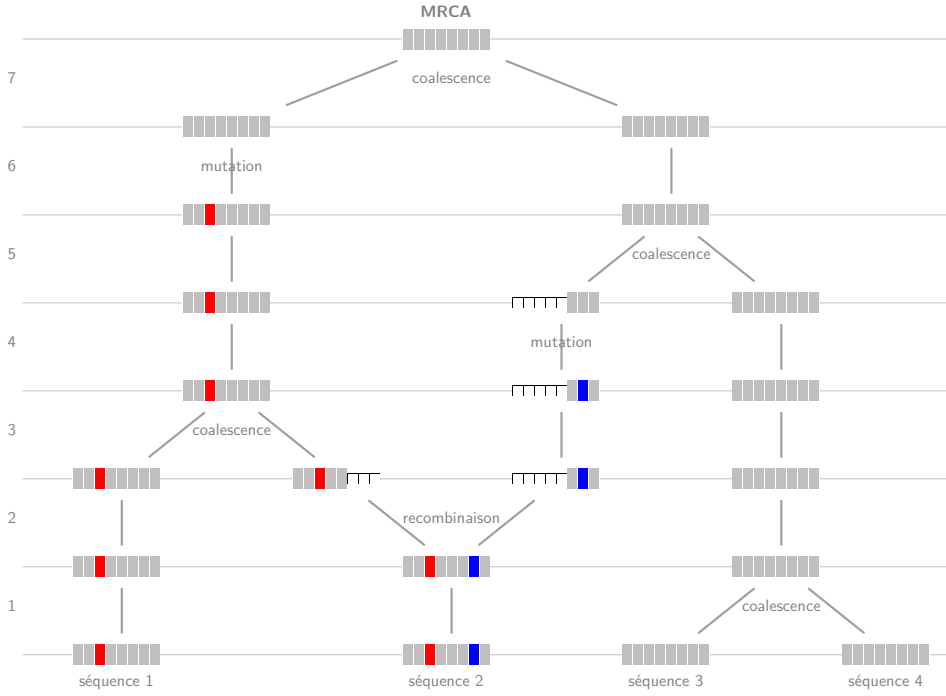


Figure 1.6 Arbre de recombinaison ancestral. Exemple d'un ARG possible pour relier 4 séquences de 8 SNPs (en bas) à leur MRCA. Sept évènements sont utilisés, soient 4 coalescences, 2 mutations et 1 recombinaison. Les allèles ancestraux primitifs sont gris, les allèles ancestraux mutés en couleurs et les non ancestraux en blanc.

ment MapARG utilise l'échantillonnage pondéré (*importance sampling*) pour parvenir à ses fins. Par contre, puisque le taux de recombinaison $\frac{n\rho}{2}$ est linéaire en n et que le taux de coalescence $\binom{n}{2} = \frac{n^2-n}{2}$ est quadratique, le MRCA est assuré d'être atteint.

1.3 Cartographie génétique

La cartographie génétique a pour but d'annoter la séquence d'ADN humain, longue de plus de 3 milliards de nucléotides, c'est-à-dire de déterminer la fonction de segments précis ou même de nucléotides dans les processus biologiques normaux et/ou anormaux, tels que des maladies. MapARG, qui sera développée au chapitre II, est une méthode de cartographie génétique fine visant à localiser le plus précisément possible un nucléotide dont la mutation aurait une influence significative sur un phénotype tel qu'une maladie. Elle est basée sur un processus de coalescence, et plus spécifiquement fait usage de l'ARG tel que nous venons de le décrire.

NOTATIONS DU CHAPITRE I

N	Taille de la population.
n_e	Taille de l'échantillon.
n	Nombre de lignées restantes à un moment donné.
μ	Taux de mutation sur une lignée entre deux générations.
$\theta = 2\,\mu N$	Taux de mutation dans la population.
r	Taux de recombinaison sur une lignée entre deux générations.
$\rho = 2\,r N$	Taux de recombinaison dans la population.
T_C^n, T_M^n, T_R^n	Temps avant qu'il y ait une coalescence, une mutation ou une recombinaison, respectivement, lorsqu'il reste n lignées.
T^n	Temps avant qu'il y ait un évènement, lorsqu'il reste n lignées.

CHAPITRE II

CARTOGRAPHIE GÉNÉTIQUE VIA UN PROCESSUS DE COALESCENCE :

LA MÉTHODE MapARG

2.1 Liaison génétique et déséquilibre de liaison

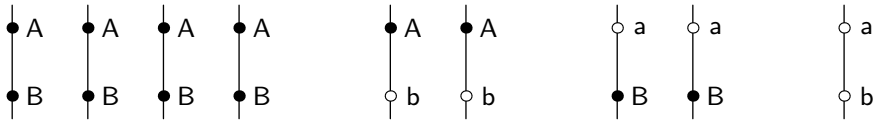
Les généticiens T. Morgan et A. Sturtevant réalisent la première carte génétique en 1913, sur le chromosome X de la drosophile. Depuis, différents types de méthodes de cartographie génétique ont été développés.

La méthode de cartographie génétique fine MapARG, présentée dans ce chapitre, repose sur le *déséquilibre de liaison* (*Linkage Disequilibrium*, ou *LD*), à travers un processus de coalescence. La liaison génétique (*genetic linkage*) est la tendance qu'ont deux marqueurs situés à proximité sur le même chromosome à être transmis ensemble lors de la méïose. Le taux de recombinaison par méïose (par génération) entre deux marqueurs situés sur deux chromosomes différents est de $\frac{1}{2}$, en raison de la recombinaison *inter*-chromosomique (page 11). Il est cependant inférieur, entre deux marqueurs situés sur le même chromosome. De fait, plus deux marqueurs sont rapprochés, plus faible est la probabilité de recombinaison *intra*-chromosomique entre eux, et donc plus forte est leur liaison génétique.

Lorsqu'une nouvelle mutation survient sur un nucléotide jusqu'ici monoallélique, le nouvel allèle est associé aux allèles présents sur le chromosome sur lequel il est apparu. Après quelques générations, les recombinaisons feront en sorte qu'il deviendra de moins en moins associé à ces allèles, particulièrement avec ceux des nucléotides éloignés sur le chromosome. Selon le modèle d'Hardy-Weinberg, si on laisse assez de temps, et si l'allèle ne subit pas l'extinction, il deviendra éventuellement en équilibre avec toute la variation nucléotidique présente dans la population. Cependant, les processus évolutifs tels que la sélection naturelle, la dérive génétique, les mutations et la migration agissent en réalité trop rapidement pour que ceci ait le temps de se produire.

Supposons deux SNPs α et β, situés sur le même chromosome. Nous noterons p_A, p_a, p_B et p_b les proportions dans la population des haplotypes porteurs des allèles A et a au SNP α, B et b au SNP β, respectivement. Ainsi, $p_A + p_a = 1 = p_B + p_b$. Nous noterons aussi p_{AB}, p_{Ab}, p_{aB} et p_{ab} les proportions des haplotypes AB, Ab, aB et ab dans la population,

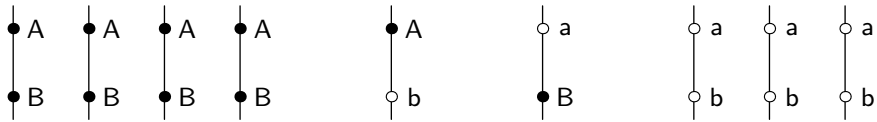
et donc $p_{AB} + p_{Ab} + p_{aB} + p_{ab} = 1$. Si tous ces allèles sont en équilibre d'Hardy-Weinberg dans la population, alors on aura que $p_{AB} = p_A p_B$, $p_{Ab} = p_A p_b$, $p_{aB} = p_a p_B$ et $p_{ab} = p_a p_b$. Dans cet exemple de 9 haplotypes :



on voit que

$$p_{AB} = \frac{4}{9} \;=\; p_A\, p_B = \left(\frac{6}{9}\right)\left(\frac{6}{9}\right) = \frac{36}{81}.$$

Cependant, si l'un des haplotypes, disons AB, se retrouve plus souvent qu'attendu dans la population, c'est-à-dire que $p_{AB} \gg p_A p_B$, alors on dit que ces allèles sont en déséquilibre de liaison, ou LD. Par exemple :



Dans ce cas, on a

$$p_{AB} = \frac{4}{9} > p_A\, p_B = \left(\frac{5}{9}\right)\left(\frac{5}{9}\right) = \frac{25}{81}.$$

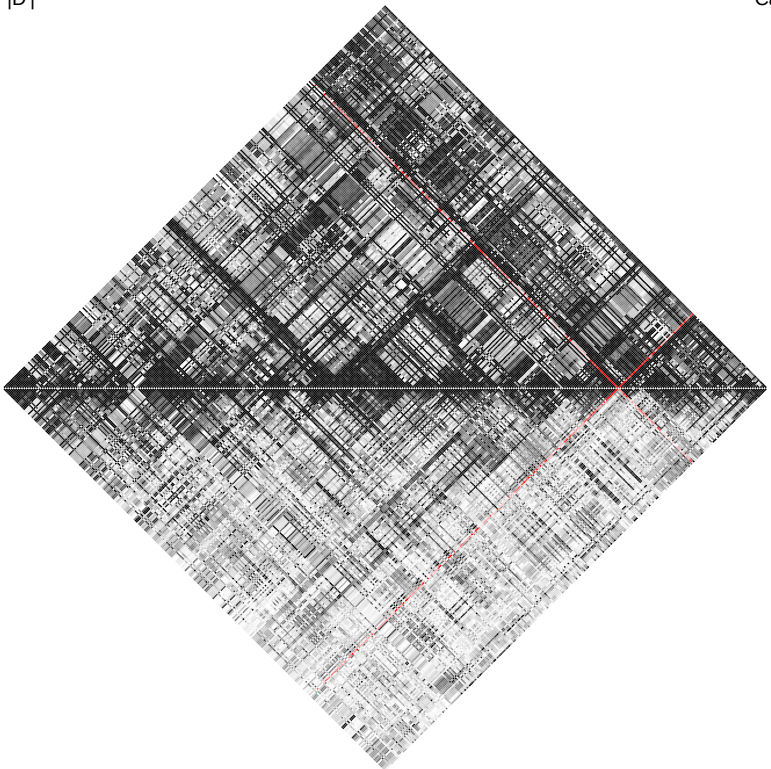
Il existe plusieurs mesures statistiques simples pour mesurer le degré de LD entre deux SNPs. L'une des plus utilisées est

$$|D'| = \left|\frac{p_{AB} - p_A p_B}{D_{\max}}\right|, \qquad \text{où} \qquad D_{\max} = \begin{cases} \min(p_A p_B,\, p_a p_b), & \text{si } p_{AB} < p_A p_B; \\ \min(p_A p_b,\, p_a p_B), & \text{si } p_{AB} > p_A p_B. \end{cases}$$

Une valeur de $|D'|$ de 0 signifie que les SNPs α et β sont en équilibre de liaison, alors que $|D'| = 1$ indique une dépendance complète. La figure 2.1 montre un exemple de $|D'|$ calculé sur un échantillon d'haplotypes. Certaines méthodes simples de cartographie génétique consistent à calculer une telle mesure de LD entre un phénotype binaire et chacun des SNPs disponibles dans un échantillon. L'hypothèse est que les SNPs se trouvant près d'une mutation influençant le phénotype seront en fort LD avec ce dernier. MapARG est aussi fondée sur la présence de LD, mais en utilisant l'information de tous les SNPs disponibles de façon non indépendante.

|D'|

Cas



Témoins

Figure 2.1 Déséquilibre de liaison mesuré par $|D'|$, entre toutes les paires de SNPs, pour les cas et les témoins d'un échantillon d'haplotypes génotypés sur 400 SNPs. Chaque petit carré représente le $|D'|$ calculé entre deux SNPs, où blanc $= 0$, noir $= 1$ et 254 teintes de gris illustrent les valeurs intermédiaires. Le rouge illustre le $|D'|$ entre le TIM (discuté plus bas) et tous les SNPs. Il s'agit donc de deux demi-matrices, l'une pour les cas (haut), l'autre pour les témoins (bas). Près de la ligne horizontale médiane, le $|D'|$ correspond à des SNPs rapprochés, alors que le $|D'|$ entre des SNPs éloignés se retrouve loin de la ligne. On peut aisément identifier des blocs de LD plus ou moins gros, le long de la ligne médiane, séparés par des points forts de recombinaison (*hotspots*). De plus, il appert évident que le LD est globalement plus important chez les cas, indiquant une divergence génétique dans cette région entre les cas et les témoins.

2.2 Objectif de la méthode

MapARG est une méthode proposée relativement récemment (Larribe, 2003 ; Larribe *et al.*, 2002). Quoiqu'elle ait déjà démontré sa capacité à localiser correctement une mutation causale dans de réelles banques de données, MapARG est encore aujourd'hui en phase de développement et d'amélioration. Comme son nom le suggère, cette méthode est basée sur la construction de graphes de recombinaison ancestraux («ARG») pour cartographier («Map») un segment d'ADN, plus spécifiquement pour localiser une *mutation influençant un caractère* ou *TIM* (*Trait Influencing Mutation*). Cette section décrit le fonctionnement de MapARG tel qu'il est aujourd'hui, en mentionnant parfois de récentes modifications. Le lecteur peut se référer aux notations mathématiques à la fin du présent chapitre (page 73) afin de se remémorer la signification d'un terme.

2.2.1 Nature des données

MapARG cherche à extraire de l'information, contenue dans les données disponibles, sur la position d'un TIM. Ces données prennent la forme d'un échantillon d'individus, certains exprimant le caractère d'intérêt, par exemple une maladie (ce sont les *cas*), et d'autres pas (ce sont les *témoins*). Tous ces individus sont génotypés pour une liste précise de marqueurs génétiques. Ces marqueurs génétiques sont des SNPs avec deux allèles possibles et sont donc binaires, comme le phénotype, qualitatif, qui ne peut prendre que deux états.

2.2.2 Idée générale

Avant d'entrer dans les détails de la méthode, il est important d'en comprendre l'idée générale. En théorie, il existe une et une seule généalogie reliant correctement tous les individus de notre échantillon à leur plus récent ancêtre commun (MRCA). Cet arbre généalogique

réel est cependant inconnu. Il est toutefois possible de construire des ARGs reliant tous les individus de l'échantillon à un MRCA et qui sont compatibles avec les données.

Comme nous l'avons vu au chapitre I, le nombre d'ARGs compatibles avec un échantillon est infini, en raison des évènements de recombinaison. Cependant, MapARG tire avantage du fait que certains ARGs sont plus probables que d'autres. Ainsi, dans un premier temps, un ARG est construit aléatoirement, suivant une certaine distribution des évènements possibles. Puis, dans un deuxième temps, la probabilité de la position du TIM est estimée pour plusieurs positions le long de la séquence. Cette probabilité correspond en fait à la probabilité d'observer les données si le TIM était à cette position et que cet ARG représentait la vraie généalogie de l'échantillon. Plusieurs ARGs (des milliers, voire des millions) sont ainsi construits aléatoirement, et une vraisemblance est obtenue pour chaque position par une moyenne pondérée des probabilités calculées à cette position. Une courbe permet ensuite de visualiser la vraisemblance de la position du TIM le long de la séquence d'ADN (figure 2.2).

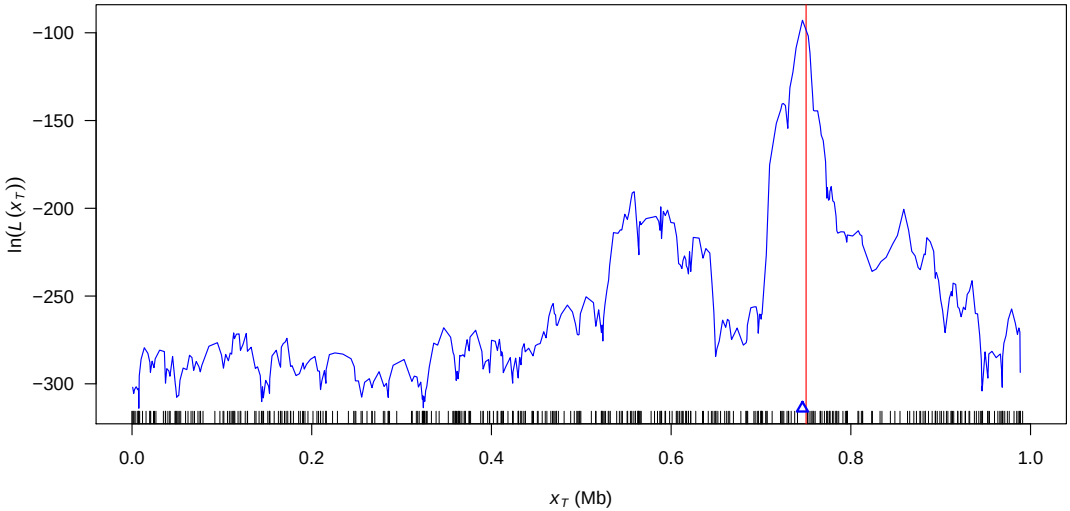


Figure 2.2 Courbe de vraisemblance obtenue par la méthode MapARG (en bleu). La position en Mb de chacun des SNPs est dénotée par un court trait vertical le long de la séquence (axe des abscisses), alors que le *log* de la vraisemblance obtenue le long de la séquence est projeté sur l'axe des ordonnées. La position réelle du TIM recherché est dénotée par une droite verticale rouge alors qu'un petit triangle bleu indique sa position estimée par MapARG, soit celle du maximum de vraisemblance. Dans cet exemple la position estimée tombe très près de la position réelle du TIM. *Détails techniques :* ce résultat a été obtenu avec un échantillon simulé de 100 cas et 100 témoins, sur une séquence de 400 SNPs étalés sur 1 Mb, à l'aide de fenêtres de 6 SNPs incrémentées de 1.

≪ < > ≫

2.2.3 Vraisemblance

Soit H_0 un échantillon d'haplotypes de L SNPs de positions connues sur un segment d'ADN de longueur R. Nous noterons x_s la position du SNP s, en Mb, où $s = 1, \ldots, L$ tel que $0 = x_1 < x_2 < \ldots < x_L = R$, et x_T notera la position du TIM. La méthode consiste conséquemment à estimer la vraisemblance

$$L(x_T) \equiv Q_{x_T}(H_0) \equiv Q(H_0|x_T) , \qquad x_T \in [x_1,x_L] .$$

Dans le but d'alléger la notation, Q_{x_T} sera parfois noté Q.

2.3 Échantillonnage pondéré

La construction d'un ARG implique le passage de l'état H_0, (l'échantillon d'haplotypes observés) à l'état H_{τ^*} (le MRCA) en passant par plusieurs états $H_1, \ldots, H_\tau, \ldots, H_{\tau^*-1}$, où un état H_τ correspond aux haplotypes restants après le τ^e évènement (de coalescence,

mutation ou recombinaison). Une fois un ARG construit, on peut calculer sa probabilité, conditionnellement au paramètre x_T, en redescendant, à partir du MRCA, jusqu'à H_0. Comme il s'agit d'un processus markovien et que la probabilité d'un état ne dépend du passé que par l'état précédent, on peut obtenir l'équation de récurrence :

$$\begin{aligned} Q_{x_T}(H_0,\ldots,H_{\tau^*}) \equiv Q(H_0,\ldots,H_{\tau^*}) &= Q(H_0|H_1)\; Q(H_1,\ldots,H_{\tau^*}) \\ &= Q(H_0|H_1)\; Q(H_1|H_2)\; Q(H_2,\ldots,H_{\tau^*}) \\ &= Q(H_0|H_1)\; Q(H_1|H_2)\; Q(H_2|H_3)\; Q(H_3,\ldots,H_{\tau^*}) \\ &\vdots \\ &= \left(\prod_{\tau=0}^{\tau^*-1} Q(H_\tau|H_{\tau+1})\right)\; Q(H_{\tau^*}). \end{aligned}$$

Puisque l'on suppose que le MRCA est connu et unique (allèle 0 à tous les SNPs), on a que $Q(H_{\tau^*}) = 1$ et donc

$$Q_{x_T}(H_0, \ldots, H_{\tau^*}) = \prod_{\tau=0}^{\tau^*-1} Q_{x_T}(H_\tau | H_{\tau+1}).$$

En supposant fini le nombre de généalogies compatibles avec H_0, on peut calculer la probabilité d'un ARG en considérant tous les ARGs possibles. Ainsi, en sommant sur tous les états possibles une étape en remontant dans le temps, on a

$$Q(H_\tau) = \sum_{H_{\tau+1}} Q(H_\tau | H_{\tau+1})\, Q(H_{\tau+1}), \tag{2.1}$$

et on obtient l'équation de récurrence :

$$
\begin{aligned}
Q_{x_T}(H_0) \equiv Q(H_0) &= \sum_{H_1} Q(H_0|H_1)\, Q(H_1) \\
&= \sum_{H_1} \left(Q(H_0|H_1) \sum_{H_2} Q(H_1|H_2)\, Q(H_2) \right) \\
&= \sum_{H_1} \left(Q(H_0|H_1) \sum_{H_2} \left(Q(H_1|H_2) \sum_{H_3} Q(H_2|H_3)\, Q(H_3) \right) \right) \\
&\quad \vdots \\
&= \sum_{H_1} \left(Q(H_0|H_1) \sum_{H_2} \left(Q(H_1|H_2) \ldots \sum_{H_{\tau^*}} Q(H_{\tau^*-1}|H_{\tau^*})\, Q(H_{\tau^*}) \right) \right) \\
&= \sum_{H_1} \left(\sum_{H_2} \left(\ldots \sum_{H_{\tau^*-1}} Q(H_0|H_1)\, Q(H_1|H_2) \ldots Q(H_{\tau^*-1}|H_{\tau^*}) \right) \right) \\
&= \sum_{H_1, \ldots, H_{\tau^*-1}} \left(\prod_{\tau=0}^{\tau^*-1} Q_{x_T}(H_\tau|H_{\tau+1}) \right).
\end{aligned}
$$

En connaissant tous les états possibles, la vraisemblance de x_T pourrait ainsi être calculée. Ce calcul est cependant irréalisable en raison de l'espace infini sur les ARGs. L'échantillonnage pondéré nous permet néanmoins d'utiliser une distribution P_{x_T} pour générer un nombre fini d'ARGs. L'espérance sur cette distribution nous donnera alors une estimation de la vraisemblance recherchée. L'équation 2.1 peut être réécrite sous la forme :

$$\begin{aligned} Q(H_\tau) &= \sum_{H_{\tau+1}} Q(H_\tau|H_{\tau+1})\, Q(H_{\tau+1}) \\ &= \sum_{H_{\tau+1}} Q(H_\tau|H_{\tau+1})\, \frac{P(H_{\tau+1}|H_\tau)}{P(H_{\tau+1}|H_\tau)}\, Q(H_{\tau+1}) \\ &= \sum_{H_{\tau+1}} h(H_\tau|H_{\tau+1})\, P(H_{\tau+1}|H_\tau)\, Q(H_{\tau+1}), \end{aligned}$$

où

$$h(H_\tau|H_{\tau+1}) = \frac{Q(H_\tau|H_{\tau+1})}{P(H_{\tau+1}|H_\tau)}$$

et donc

$$Q(H_\tau|H_{\tau+1}) = h(H_\tau|H_{\tau+1})\ P(H_{\tau+1}|H_\tau).$$

La vraisemblance recherchée peut alors être réécrite sous la forme d'une espérance sur la distribution P_{x_T} :

$$\begin{aligned}
Q_{x_T}(H_0) \equiv Q(H_0) &= \sum_{H_1,\ldots,H_{\tau^*-1}} \left(\prod_{\tau=0}^{\tau^*-1} h(H_\tau|H_{\tau+1})\ P(H_{\tau+1}|H_\tau) \right) \\
&= \sum_{H_1,\ldots,H_{\tau^*-1}} \left(\prod_{\tau=0}^{\tau^*-1} h(H_\tau|H_{\tau+1}) \right) \left(\prod_{\tau=0}^{\tau^*-1} P(H_{\tau+1}|H_\tau) \right) \\
&= \mathrm{E}_{P_{x_T}} \left[\prod_{\tau=0}^{\tau^*-1} h(H_\tau|H_{\tau+1}) \right] \\
&= \mathrm{E}_{P_{x_T}} \left[\prod_{\tau=0}^{\tau^*-1} \frac{Q(H_\tau|H_{\tau+1})}{P(H_{\tau+1}|H_\tau)} \right].
\end{aligned}$$

Quoiqu'explicitement incalculable, cette espérance nous permet cependant d'estimer $L(x_T)$, la vraisemblance de la position du TIM, par une moyenne sur un certain nombre K d'ARGs

générés selon la distribution P_{x_T} :

$$\hat{L}(x_T) \;=\; \hat{Q}_{x_T}(H_0) \;=\; \frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{P_{x_T}(H_{\tau+1}|H_\tau)}\right). \tag{2.2}$$

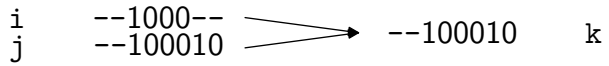
Afin d'estimer cette vraisemblance, il est nécessaire de connaître les deux distributions Q (chronologique) et P (en remontant dans le temps).

2.4 Coalescences, mutations et recombinaisons

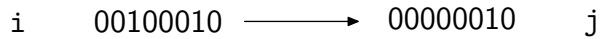
Rappelons que lors de la construction d'un ARG, chaque étape est constituée d'un évènement parmi les trois types possibles :

C_{ij}^k **Coalescence de séquences de types i et j compatibles en une séquence parentale de type k.** Deux séquences de types i et j peuvent coalescer en une séquence parentale de type k si les SNPs ancestraux (0 ou 1) qu'elles

possèdent en commun portent les mêmes allèles. Comme dans cet exemple, toute l'information ancestrale est alors conservée :



$M_i^j(s)$ **Mutation au SNP s d'une séquence de type i en une séquence parentale de type j.** Rappelons que les mutations sont modélisées selon le modèle de mutations à sites infinis, et qu'elles ne peuvent donc se produire qu'une seule fois sur un SNP donné. Conséquemment, un évènement de mutation, en remontant dans le temps, ne peut survenir que s'il ne reste plus qu'une seule séquence portant l'allèle muté (1) sur ce SNP s. Dans cet exemple, on suppose que la séquence i est la dernière à porter l'allèle muté au SNP $s = 3$:



$R_i^{jk}(s)$ **Recombinaison, dans l'intervalle s, d'une séquence de type i en deux séquences parentales de types j et k.** Une séquence entièrement non ancestrale (-) n'apportant aucune information sur la généalogie construite, seules seront possibles les recombinaisons dans un intervalle s où chacune des deux séquences parentales résultantes contient au moins 1 SNP ancestral (0 ou 1). Ici, $s = 4$:

2.4.1 Taux et probabilités des évènements

Après chaque évènement, lors de la construction d'un ARG, les probabilités des différents évènements devront être mises à jour en fonction des séquences restantes. Tout d'abord, le

taux de coalescence ne dépendra que du nombre n de séquences restantes et, comme vu à la section 1.2.3.1, sera :

$$\pi_C \;=\; \binom{n}{2} \;=\; \frac{n}{2}(n-1).$$

Cependant, les taux de mutation et de recombinaison dépendront également du nombre de SNPs ancestraux (0 ou 1) présents dans les séquences. Soient μ_s le taux de mutation par génération du SNP s, pour $s = 1, \ldots, L$ et $\theta_s = 2\,\mu_s N$ le taux à l'échelle de coalescence. Le taux de mutation sur l'ensemble des SNPs est donc $\theta = \sum_s \theta_s$. Nous noterons n^s le nombre de séquences restantes contenant le SNP s sous forme ancestrale (0 ou 1). Ainsi, le taux de mutation sera donné par :

$$\pi_M \;=\; \frac{n}{2}(\alpha\,\theta) \;=\; \frac{1}{n}\sum_s n^s\theta_s\;, \qquad \text{où} \qquad \alpha \;=\; \frac{2}{n\,\theta}\,\frac{1}{n}\sum_s n^s\theta_s.$$

De la même manière, le taux de recombinaison à l'échelle de coalescence entre les SNPs s et $s+1$ est donné par $\rho_s = 2\,r_s N$, où r_s est le taux par génération, pour $s = 1, \ldots, L-1$.

Ainsi, le taux de recombinaison sur toute la région entre les SNPs 1 et L est $\rho = \sum_s \rho_s$. Nous noterons $n^{|s|}$ le nombre de séquences restantes contenant au moins 1 SNP de chaque côté de l'intervalle s sous forme ancestrale (0 ou 1). Ainsi, le taux de recombinaison sera :

$$\pi_R \;=\; \frac{n}{2}(\beta\,\rho) \;=\; \frac{1}{n}\sum_s n^{|s|}\rho_s\,, \qquad \text{où} \qquad \beta \;=\; \frac{2}{n\,\rho}\,\frac{1}{n}\sum_s n^{|s|}\rho_s.$$

Donc, en remontant dans le temps, les probabilités que le prochain évènement τ soit une coalescence, une mutation ou une recombinaison seront données par :

$$P_\tau(C) \;=\; \frac{\pi_C}{\pi_C+\pi_M+\pi_R} \;=\; \frac{n-1}{n-1+\alpha\,\theta+\beta\,\rho};$$

$$P_\tau(M) \;=\; \frac{\pi_M}{\pi_C+\pi_M+\pi_R} \;=\; \frac{\alpha\,\theta}{n-1+\alpha\,\theta+\beta\,\rho};$$

$$P_\tau(R) \;=\; \frac{\pi_R}{\pi_C+\pi_M+\pi_R} \;=\; \frac{\beta\,\rho}{n-1+\alpha\,\theta+\beta\,\rho}.$$

2.4.2 Distribution Q

Maintenant que nous avons la distribution sur le type du τ^e évènement, nous allons pouvoir obtenir la distribution $Q_{x_T}(H_\tau|H_{\tau+1})$ sur tous les évènements possibles des 3 types. Nous noterons n_i le nombre de séquences de type i restantes, et donc $n = \sum_i n_i$. Si le τ^e évènement est une coalescence, il restera $n-1$ séquences, à l'état $H_{\tau+1}$, pouvant résulter d'une coalescence de deux séquences de l'état H_τ. De plus, s'il s'agit de la coalescence de séquences de types i et j en une séquence de type k, il restera $n_k + 1 - \delta_{ik} - \delta_{jk}$ séquences de type k, où $\delta_{ik} = 1$ si $i = k$ et 0 sinon (et $\delta_{jk} = 1$ si $j = k$ et 0 sinon). Si $i = j = k$, il restera donc $n_k - 1$ séquences de ce type à l'état $H_{\tau+1}$. Ainsi, si le τ^e évènement est une coalescence, il s'agira d'une coalescence d'une séquence de type i et d'une séquence de type j en une séquence de type k avec probabilité

$$P_\tau(C_{ij}^k|C) \;=\; \frac{\binom{n_k+1-\delta_{ik}-\delta_{jk}}{1}}{\binom{n-1}{1}} \;=\; \frac{n_k+1-\delta_{ik}-\delta_{jk}}{n-1}.$$

Si le τ^e évènement est une mutation, d'une séquence de type i en une séquence de type j, il restera, à l'état $H_{\tau+1}$, n_i-1 séquences de type i, n_j+1 séquences de type j et donc n séquences. Puisque la probabilité que la mutation survienne au SNP s est de $\frac{\theta_s}{\alpha\,\theta}$, la probabilité d'une mutation $M_i^j(s)$ sera donc donnée par

$$P_\tau(M_i^j(s)|M) \;=\; \frac{\theta_s}{\alpha\,\theta}\frac{\binom{n_j+1}{1}}{n} \;=\; \frac{\theta_s(n_j+1)}{\alpha\,\theta\,n}.$$

Enfin si le τ^e évènement est une recombinaison, d'une séquence de type i en séquences de types j et k, il y aura $n_i - 1$ séquences de type i, $n_j + 1$ séquences de type j, $n_k + 1$ séquences de type k et donc $n + 1$ séquences, à l'état $H_{\tau+1}$. Il y a donc $\binom{n_j+1}{1}\binom{n_k+1}{1}$ combinaisons possibles d'une séquence partielle de gauche de type j et d'une séquence partielle de droite de type k. Cependant, puisque chaque type de séquence pourrait potentiellement servir de séquence partielle gauche ou droite, il y a $2\binom{n+1}{1}$ combinaisons possibles de 2 séquences dans $H_{\tau+1}$. Comme la probabilité que la recombinaison survienne dans l'intervalle s est de $\frac{\rho_s}{\beta\,\rho}$, la probabilité d'une recombinaison $R_i^{jk}(s)$ sera donc donnée par

$$P_\tau(R_i^{jk}(s)|R) \;=\; \frac{\rho_s}{\beta\,\rho}\frac{\binom{n_j+1}{1}\binom{n_k+1}{1}}{2\binom{n+1}{1}} \;=\; \frac{\rho_s(n_j+1)(n_k+1)}{\beta\,\rho\,n(n+1)}.$$

En notant $H_\tau + E$ l'état $H_{\tau+1}$ résultant d'un évènement E à partir de l'état H_τ, on a la distribution chronologique Q :

$$Q_{x_T}(H_\tau | H_\tau + E) = \begin{cases} P_\tau(C_{ij}^k | C)\ P_\tau(C), & \text{si } E = C_{ij}^k; \\ P_\tau(M_i^j(s) | M)\ P_\tau(M), & \text{si } E = M_i^j(s); \\ P_\tau(R_i^{jk}(s) | R)\ P_\tau(R), & \text{si } E = R_i^{jk}(s); \end{cases}$$

$$= \begin{cases} \frac{n_k + 1 - \delta_{ik} - \delta_{jk}}{n - 1 + \alpha\,\theta + \beta\,\rho}, & \text{si } E = C_{ij}^k; \\ \frac{\theta_s (n_j + 1)}{n(n - 1 + \alpha\,\theta + \beta\,\rho)}, & \text{si } E = M_i^j(s); \\ \frac{\rho_s (n_j + 1)(n_k + 1)}{n(n+1)(n - 1 + \alpha\,\theta + \beta\,\rho)}, & \text{si } E = R_i^{jk}(s). \end{cases}$$

Ainsi, l'équation 2.1 de récurrence devient (similairement à celle proposée dans Larribe *et al.*, 2002) :

$$Q_{x_T}(H_\tau) = \sum_{i,j_{(s_i=s_j,\forall s_i,s_j\in\{0,1\})}} \frac{n_k+1-\delta_{ik}-\delta_{jk}}{n-1+\alpha\,\theta+\beta\,\rho}\,Q(H_\tau+C_{ij}^k)$$

$$+\sum_{i\ (n_i=1)}\ \sum_{s\ \binom{s_i=1;}{s_k\neq 1,\forall k\neq i}} \frac{\theta_s(n_j+1)}{n(n-1+\alpha\,\theta+\beta\,\rho)}\,Q(H_\tau+M_i^j(s))$$

$$+\sum_{i}\ \sum_{s\ \binom{s_i,s_j\in\{0,1\};}{(s+1)_i,(s+1)_j\in\{0,1\}}} \frac{\rho_s(n_j+1)(n_k+1)}{n(n+1)(n-1+\alpha\,\theta+\beta\,\rho)}\,Q(H_\tau+R_i^{jk}(s)),$$

où s_i dénote l'allèle d'une séquence de type i au SNP s.

2.4.3 Distribution P

La vraisemblance est estimée en utilisant une distribution proposée pour construire les graphes, distribution adaptée de Fearnhead et Donnelly, 2001, et récemment implantée dans MapARG par Descary (2012). Cette distribution est très importante pour la méthode

MapArg, car les inférences de la méthode dépendent en bonne partie de la qualité de cette distribution à construire des graphes qui décrivent des histoires pouvant adéquatement décrire l'évolution de l'échantillon. De nombreux travaux ont été présentés ces 20 dernières années sur cette question (Griffiths et Tavaré, 1994a,1994b,1994c ; Kuhner *et al.*, 1995 ; Stephens et Donnelly, 2000). Seules les grandes lignes étant ici présentées, le lecteur désirant plus de détails sur cette implantation peut se référer à ces ouvrages. Cette distribution proposée P est donnée par

$$P_{x_T}(H_{\tau+1}|H_\tau) \;=\; Q_{x_T}(H_\tau|H_{\tau+1})\;\frac{\phi(H_{\tau+1})}{\phi(H_\tau)},$$

où

$$\frac{\phi(H_\tau+E)}{\phi(H_\tau)} \;=\; \begin{cases} \frac{\phi(k|H_\tau-i-j)}{\phi(i|H_\tau-i)\;\phi(j|H_\tau-i-j)}, & \text{si } E=C_{ij}^k; \\ \frac{\phi(j|H_\tau-i)}{\phi(i|H_\tau-i)}, & \text{si } E=M_i^j(s); \\ \frac{\phi(j|H_\tau-i)\;\phi(k|H_\tau-i+j)}{\phi(i|H_\tau-i)}, & \text{si } E=R_i^{jk}(s), \end{cases}$$

et $\phi(i|H_\tau - i)$ représente la probabilité de piger aléatoirement une séquence de type i parmi une population, lorsque l'on a déjà pigé les séquences présentes dans $H_\tau - i$.

Cette distribution nous permet finalement d'estimer la vraisemblance de la position du TIM par l'équation 2.2 (page 56), qui devient :

$$\begin{aligned}
\hat{L}(x_T) &= \frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{P_{x_T}(H_{\tau+1}|H_\tau)}\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{Q_{x_T}(H_\tau|H_{\tau+1})\,\frac{\phi(H_{\tau+1})}{\phi(H_\tau)}}\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{\phi(H_\tau)}{\phi(H_{\tau+1})}\right).
\end{aligned}$$

2.5 Vraisemblance composite

La taille d'un échantillon (le nombre d'haplotypes mais aussi le nombre de SNPs) est un

facteur très important dans le temps de calcul requis par MapARG. Avec la disponibilité grandissante des données génétiques, plus d'individus génotypés sur plus de SNPs permettront d'obtenir de meilleurs résultats, plus précis et plus fiables. Les capacités informatiques n'étant toutefois pas infinies, quoiqu'en constante progression, l'intégration de nouveaux modèles mathématiques visant à réduire les temps de calcul devient indispensable.

La vraisemblance composite est de plus en plus utilisée en statistique génétique (Larribe et Fearnhead, 2011) en raison de la quantité croissante de données disponibles et de leur dépendance évidente du point de vue génétique. Elle fut récemment implantée dans MapARG (Larribe et Lessard, 2008), réduisant considérablement le temps de calcul et permettant ainsi l'utilisation de beaucoup de SNPs et d'échantillons de grande taille.

La vraisemblance composite est appliquée dans MapARG à travers l'utilisation partielle des SNPs disponibles. L'ensemble des SNPs de l'échantillon est divisé en $L - d + 1$ fenêtres consécutives de d SNPs chacune (figure 2.3). Des ARGs sont construits avec les haplotypes

partiels d'une fenêtre, et la vraisemblance $L(x_T|x_s < x_T < x_{s+1})$ est estimée par toutes les fenêtres g englobant l'intervalle s, soit l'ensemble G_s (rappelons que l'intervalle s est situé entre les SNPs s et $s+1$).

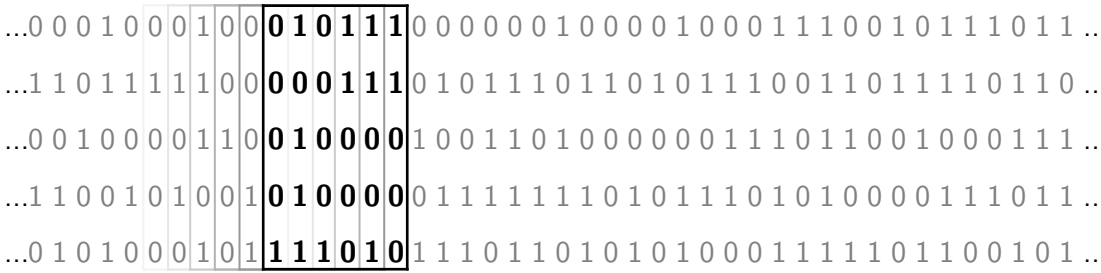


Figure 2.3 Vraisemblance composite dans MapARG. Exemple d'une fenêtre de $d = 6$ SNPs déplacée par incrément de 1 SNP, sur un échantillon de 5 haplotypes. Les 5 haplotypes partiels de la fenêtre courante sont en noir.

Nous noterons $L_{s,g}(x_T)$ la vraisemblance marginale de la position x_T du TIM dans l'intervalle s donnée par les haplotypes de la fenêtre g. La vraisemblance composite (*composite*

likelihood, CL), pondérée uniformément par le nombre $|G_s|$ de fenêtres couvrant un intervalle s, est alors donnée par (Larribe et Lessard, 2008) :

$$CL(x_T) \;=\; \prod_{s=1}^{L-1}\left(\prod_{g\in G_s} L_{s,g}(x_T)\right)^{\frac{1}{|G_s|}},$$

et on a donc

$$\hat{CL}(x_T) \;=\; \prod_{s=1}^{L-1}\left(\prod_{g\in G_s}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{\phi(H_\tau)}{\phi(H_{\tau+1})}\right)\right)\right)^{\frac{1}{|G_s|}}.$$

Il est à noter que, à l'exception des $d-2$ premiers et derniers intervalles, les $L-2d+3$ autres intervalles seront englobés par $d-1$ fenêtres, d'où l'importance de pondérer la vraisemblance. De plus, si $d = L$, alors $\hat{CL}(x_T) = \hat{L}(x_T)$.

2.6 Algorithme de MapARG

Afin de bien saisir la structure de la méthode MapARG que nous venons de présenter, en

voici les grandes étapes :

I. Choisir l'ensemble des positions x_T pour lesquelles $CL(x_T)$ sera évaluée ;

II. Pour chacune des $L - d + 1$ fenêtres couvrant l'ensemble des SNPs de l'échantillon :

    Pour chacun des $d - 1$ intervalles situés dans la fenêtre :

        Pour chacun des K graphes à construire :

            Pour chaque étape τ du graphe, tant que le MRCA n'est pas atteint :

            i. Calculer $\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{P_{x_T}(H_{\tau+1}|H_\tau)} = \frac{\phi(H_\tau)}{\phi(H_{\tau+1})}$;

            ii. Mettre à jour $Q_{x_T}(H_\tau)$ et $P_{x_T}(H_{\tau+1})$;

            iii. Générer un évènement selon $P_{x_T}(H_{\tau+1})$;

            iv. Mettre à jour $H_{\tau+1}$;

III. Pour chaque position x_T :

    Calculer $\hat{CL}(x_T)$;

IV. $\hat{x}_T$ correspond au maximum de $\hat{CL}(x_T)$.

2.7 Inférence sur l'allèle du TIM

La méthode MapARG décrite dans ce chapitre cherche à connaître la position du TIM sur l'ADN (x_T), sous l'hypothèse que l'allèle du TIM est connu pour chacun des haplotypes. C'est à partir du phénotype de chaque haplotype que cette information est inférée. Or, jusqu'à maintenant, une maladie récessive et rare était assumée, et l'allèle du TIM était inféré directement du phénotype. Cependant, il est plutôt rare que l'allèle d'un TIM ne détermine directement le phénotype en question. C'est sur ce problème que l'on se penche au chapitre III.

NOTATIONS DU CHAPITRE II

C_{ij}^{k}	Coalescence de séquences de types i et j compatibles en une séquence de type k.
$M_{i}^{j}(s)$	Mutation au marqueur s d'une séquence de types i en une séquence parentale de type j.
$R_{i}^{jk}(s)$	Recombinaison, dans l'intervalle s d'une séquence de types i en deux séquences parentales de types j et k.
H_0	Ensemble des haplotypes observés.
H_τ	Ensemble des haplotypes restants après le τ^e évènement.
H_{τ^*}	MRCA.
L	Nombre de marqueurs.
π_C	Taux de coalescence.
π_M	Taux de mutation.

π_R	Taux de recombinaison.
x_i	Position du marqueur i (en Mb).
x_T	Position du TIM (en Mb).

CHAPITRE III

ESTIMATION DE L'ALLÈLE DU TIM

3.1 Problématique

Afin de localiser une mutation (TIM), la méthode de cartographie MapARG présentée au chapitre II requiert un échantillon d'haplotypes, ainsi que l'allèle de ce TIM associé à chacun de ces haplotypes. Or, comme discuté au chapitre I, les humains sont diploïdes. Pour une séquence d'ADN, un individu possède deux haplotypes, l'un hérité de sa mère, l'autre de son père. De plus, pour un TIM cherché, il en portera deux allèles, l'un situé sur son haplotype maternel, l'autre sur son haplotype paternel. Les données d'un individu dans un échantillon utilisable par MapARG doivent donc être sous la forme d'un *diplotype*, incluant les allèles au TIM :

	SNPs	**TIM**
haplotype 1	10110010	0
haplotype 2	00100010	1

Cependant, les données sont le plus souvent obtenues sous la forme de *génotypes*, où on ne

peut distinguer sur quel haplotype se trouve chacun des deux allèles des SNPs pour lesquels l'individu est *hétérozygote*, c'est-à-dire pour lesquels il possède deux allèles différents. De plus, le seul phénotype de l'individu ne nous informe pas directement sur les deux allèles qu'il possède au TIM, encore moins sur leur répartition dans ses deux haplotypes :

SNPs								**phénotype**
10	00	11	10	00	00	11	00	1

Dans le passé, MapARG supposait un modèle récessif où le TIM était rare, ce qui fait que les allèles au TIM pouvaient être inférés à partir du phénotype, mais ce modèle est très limitant. De ce fait, la figure 3.1 montre un exemple de résultat de vraisemblance obtenue avec MapARG sur un échantillon dont le modèle n'est pas récessif. En haut, la vraisemblance est obtenue en assumant que le modèle est récessif, et en bas la méthode présentée dans ce chapitre est d'abord utilisée pour estimer les allèles au TIM de chaque individu.

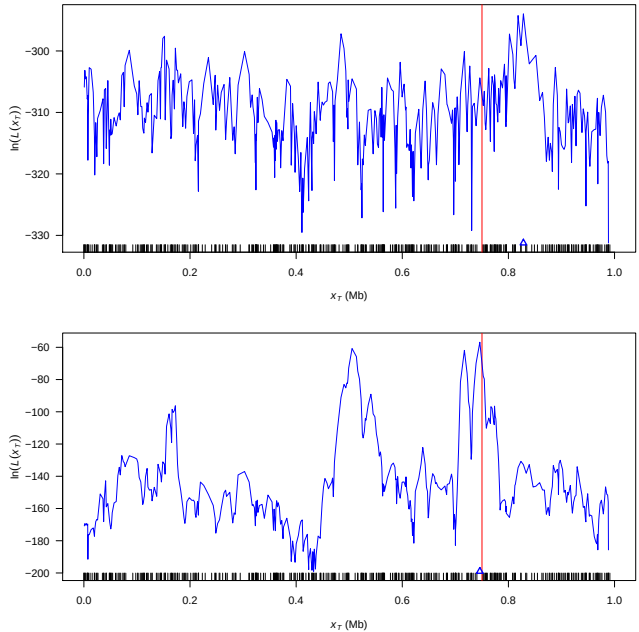


Figure 3.1 Exemple de résultats obtenus avec MapARG sur un échantillon non récessif. La courbe de vraisemblance du haut est obtenue en assumant que le modèle est récessif, alors qu'en bas la méthode présentée dans ce chapitre est d'abord utilisée pour estimer les allèles au TIM de chaque individu. On peut noter la différence dans l'échelle des courbes de vraisemblance, ainsi que dans leur variabilité. De plus, la position estimée du TIM, erronée avec la version *naïve* de MapARG, se trouve rétablie par l'estimation préalable des allèles au TIM à l'aide de la méthode. *Détails techniques :* ces deux résultats ont été obtenus avec le même échantillon de 100 cas et 100 témoins dont les phénotypes furent simulés avec le modèle de pénétrance $F = \{0{,}01,\ 0{,}02,\ 0{,}1\}$. Séquence de 400 SNPs étalés sur 1 Mb ; fenêtres de 6 SNPs incrémentées de 1.

Plusieurs méthodes ont été développées dans les dernières années pour estimer les haplotypes à partir des génotypes d'un échantillon, mais aucune ne permet d'estimer les allèles d'un TIM à partir du phénotype. Pour cette raison, un algorithme EM qui estime à la fois les diplotypes et les allèles du TIM, à partir des génotypes et des phénotypes, a récemment été développé et implanté dans MapARG par Boucher, 2009. Nous nous attarderons à tester l'efficacité de cette méthode à estimer les allèles du TIM, indépendamment de l'estimation des diplotypes. Ainsi, nous assumerons que les données sont déjà *phasées*, c'est-à-dire que les diplotypes sont connus, et l'algorithme s'en trouvera quelque peu simplifié. Le lecteur désirant connaître les détails de la méthode incluant l'estimation des diplotypes peut se référer au *chapitre IV* de l'ouvrage de Boucher. Nous nous concentrerons à développer ici une version de l'algorithme EM dans l'unique but d'estimer les allèles au TIM, car ceci est précisément un des objectifs du présent travail.

3.2 Méthode

3.2.1 Vraisemblance et étape M

Nous noterons $\delta \in \{0,1\}$ l'allèle au TIM d'un haplotype. De plus, pour un individu, nous noterons $\phi \in \{0,1\}$ son phénotype et $T = \delta_1\delta_2$ son diplotype au TIM, où δ_1, $\delta_2 \in \{0,1\}$ représentent l'allèle au TIM sur ses haplotypes maternel et paternel, respectivement. Nous noterons aussi V_0 et V_1 les distributions des haplotypes porteurs, respectivement, des allèles 0 et 1 au TIM. Ainsi, $V_\delta(h)$ dénotera la fréquence des haplotypes de type h parmi ceux qui portent l'allèle δ au TIM. Un *modèle de pénétrance* $F = \{f_0, f_1, f_2\}$ sera associé au TIM, où f_i est la probabilité pour un individu d'avoir le phénotype $\phi = 1$ s'il porte i allèles 1 (mutant) au TIM. Enfin, p et f dénoteront respectivement les fréquences, dans la population, des haplotypes portant l'allèle $\delta = 1$ au TIM et des individus diploïdes ayant le phénotype $\phi = 1$. Le lecteur pourra se référer à la page 157 pour un rappel des notations

du présent chapitre.

En supposant la population en équilibre d'Hardy-Weinberg et que l'on connait p, f et F, la distribution des diplotypes au TIM par rapport au phénotype peut aisément être calculée (tableau 3.1). La prévalence du phénotype $\phi = 1$, soit f, est habituellement connue. Si on ne connait pas p, soit la fréquence de l'allèle 1 au TIM dans la population, elle peut être obtenue de f et F (calcul détaillé à la section 4.2.1.1). Cependant, le modèle de pénétrance F étant souvent inconnu, une méthode pour l'estimer sera présentée au chapitre IV.

Tableau 3.1 Distribution des allèles au TIM dans la population

		Phénotype		**Total**
		$\phi = 0$	$\phi = 1$	
Diplotype au TIM	$T = 00$	$(1 - f_0)\,(1 - p)^2$	$f_0\,(1 - p)^2$	$(1 - p)^2$
	$T = 01$	$(1 - f_1)\,p(1 - p)$	$f_1\,p(1 - p)$	$p(1 - p)$
	$T = 10$	$(1 - f_1)\,p(1 - p)$	$f_1\,p(1 - p)$	$p(1 - p)$
	$T = 11$	$(1 - f_2)\,p^2$	$f_2\,p^2$	p^2
	Total	$1 - f$	f	1

Soit D un échantillon aléatoire simple de diplotypes tirés de la population et auxquels sont associés les phénotypes Φ. Le même échantillon de diplotypes, mais incluant les TIMs (que l'on ne connait pas), sera dénoté par D^*. Connaissant le diplotype $d = [h_1,h_2]$ d'un individu, où h_1 et h_2 dénotent respectivement ses haplotypes maternel et paternel, la probabilité du

diplotype T au TIM d'un individu est donnée par

$$\begin{aligned} P(T = \delta_1\delta_2 \mid d = [h_1,h_2], V_0,V_1) &= P(h_1|\delta_1)\, P(h_2|\delta_2)\, P(T = \delta_1\delta_2) \\ &= V_{\delta_1}(h_1)\, V_{\delta_2}(h_2)\, P(T = \delta_1\delta_2). \end{aligned}$$

Puisque le phénotype ne dépend du diplotype que par les allèles au TIM (à travers le modèle de pénétrance), on peut déduire la probabilité conjointe du diplotype au TIM et du phénotype :

$$\begin{aligned} P(T = \delta_1\delta_2,\ \phi \mid d = [h_1,h_2], V_0,V_1) &= P(\phi \mid T = \delta_1\delta_2, d, V_0,V_1)\, P(T = \delta_1\delta_2 \mid d, V_0,V_1) \\ &= P(\phi \mid T = \delta_1\delta_2)\, P(T = \delta_1\delta_2)\, V_{\delta_1}(h_1)\, V_{\delta_2}(h_2) \\ &= P(\phi,\ T = \delta_1\delta_2)\, V_{\delta_1}(h_1)\, V_{\delta_2}(h_2). \qquad (3.1) \end{aligned}$$

Les probabilités conjointes $P(\phi,\ T = \delta_1\delta_2)$ étant connues (tableau 3.1), on obtient :

$$
\begin{aligned}
P(T = 00,\ \phi = 0 \mid d,\ V_0,\ V_1) &= V_0(h_1)\, V_0(h_2)\ (1 - f_0)\, (1 - p)^2;\\
P(T = 01,\ \phi = 0 \mid d,\ V_0,\ V_1) &= V_0(h_1)\, V_1(h_2)\ (1 - f_1)\, p(1 - p);\\
P(T = 10,\ \phi = 0 \mid d,\ V_0,\ V_1) &= V_1(h_1)\, V_0(h_2)\ (1 - f_1)\, p(1 - p);\\
P(T = 11,\ \phi = 0 \mid d,\ V_0,\ V_1) &= V_1(h_1)\, V_1(h_2)\ (1 - f_2)\, p^2;\\
P(T = 00,\ \phi = 1 \mid d,\ V_0,\ V_1) &= V_0(h_1)\, V_0(h_2)\ f_0\, (1 - p)^2;\\
P(T = 01,\ \phi = 1 \mid d,\ V_0,\ V_1) &= V_0(h_1)\, V_1(h_2)\ f_1\, p(1 - p);\\
P(T = 10,\ \phi = 1 \mid d,\ V_0,\ V_1) &= V_1(h_1)\, V_0(h_2)\ f_1\, p(1 - p);\\
P(T = 11,\ \phi = 1 \mid d,\ V_0,\ V_1) &= V_1(h_1)\, V_1(h_2)\ f_2\, p^2.
\end{aligned}
$$

Les individus étant indépendants entre eux, on obtient la probabilité conjointe par le produit des probabilités marginales, et la vraisemblance sur les données est donnée par :

$$\begin{aligned} L_c(V_0,V_1) &= P(D^*,\, \Phi \mid V_0,\, V_1) \\ &= \prod_i P(d_i^*,\, \phi_i \mid V_0,\, V_1) \\ &= \prod_i P(T_i = \delta_1^i \delta_2^i,\ \phi_i \mid V_0,\, V_1) \\ &= \prod_i P(\phi_i,\ T_i = \delta_1^i \delta_2^i)\ V_{\delta_1^i}(h_1^i)\ V_{\delta_2^i}(h_2^i), \end{aligned}$$

où d_i^* est le diplotype d'un individu i, incluant ses allèles au TIM. Comme $P(\phi_i,\ T_i = \delta_1^i \delta_2^i)$ ne dépend pas des distributions à estimer, mais seulement du modèle de pénétrance F et de la fréquence p de l'allèle 1 au TIM dans la population (tableau 3.1), la vraisemblance peut être réécrite sous la forme

$$L_c(V_0,V_1) \;=\; K(F,\, p) \prod_i V_{\delta_1^i}(h_1^i)\ V_{\delta_2^i}(h_2^i),$$

qui est aussi équivalent à

$$L_c(V_0,V_1) \;=\; K(F,\,p)\prod_h V_0(h)^{m_{h^0}}\; V_1(h)^{m_{h^1}},$$

où m_{h^δ} est le nombre d'haplotypes de type h porteurs de l'allèle δ au TIM dans D^*. Ainsi, nous nous retrouvons de nouveau avec une vraisemblance de la famille exponentielle de lois, m_{h^δ} étant la statistique exhaustive pour V_δ. En notant l'espérance des statistiques exhaustives sous la forme

$$m_{h^\delta}^{(k+1)} \;=\; \mathrm{E}\left(m_{h^\delta} \mid V_0^{(k)},\, V_1^{(k)},\, D,\, \Phi\right),$$

où k dénote l'itération de l'algorithme EM, la fonction à maximiser est alors :

$$W\left(V_0,\, V_1 \mid V_0^{(k)},\, V_1^{(k)}\right) \;=\; \sum_h \left(m_{h^0}^{(k+1)}\mathsf{ln}(V_0(h)) \;+\; m_{h^1}^{(k+1)}\mathsf{ln}(V_1(h))\right),$$

sous les contraintes $\sum_h V_0(h) \;=\; \sum_h V_1(h) \;=\; 1$. L'incorporation d'un multiplicateur de Lagrange pour chacune de ces contraintes nous donne

$$W_L\left(V_0,\, V_1 \mid V_0^{(k)},\, V_1^{(k)}\right) \;=\; \lambda_0\left(1-\sum_h V_0(h)\right) \;+\; \sum_h \left(m_{h^0}^{(k+1)} \ln(V_0(h))\right)$$
$$+\; \lambda_1\left(1-\sum_h V_1(h)\right) \;+\; \sum_h \left(m_{h^1}^{(k+1)} \ln(V_1(h))\right),$$

et on obtient, en annulant les dérivés partielles, que W_L est maximale quand

$$V_0(h) = \frac{m_{h^0}^{(k+1)}}{\lambda_0} \qquad \text{et} \qquad V_1(h) = \frac{m_{h^1}^{(k+1)}}{\lambda_1}.$$

Avec les contraintes, l'étape M de l'algorithme EM consiste donc à estimer, pour tout h,

$$V_0(h)^{(k+1)} = \frac{m_{h^0}^{(k+1)}}{m_0^{(k+1)}} \qquad \text{et} \qquad V_1(h)^{(k+1)} = \frac{m_{h^1}^{(k+1)}}{m_1^{(k+1)}}, \tag{3.2}$$

où $m_\delta^{(k+1)} = \sum_h m_{h^\delta}^{(k+1)}$ est le nombre moyen d'haplotypes porteurs de l'allèle δ au TIM après l'itération k.

3.2.2 Espérances conditionnelles et étape E

Afin de compléter une itération de l'algorithme EM, il nous faut maintenant estimer les espérances conditionnelles $m_{h^\delta}^{(k+1)} = \mathrm{E}\left(m_{h^\delta} \,\middle|\, V_0^{(k)}, V_1^{(k)}, D, \Phi\right)$. En conditionnant sur le phénotype ϕ, l'équation 3.1 devient :

$$\begin{aligned} P(T = \delta_1\delta_2,\ \phi \mid d, V_0, V_1) &= P(\phi,\ T = \delta_1\delta_2)\ V_{\delta_1}(h_1)\ V_{\delta_2}(h_2) \\ &= P(\phi)\ P(T = \delta_1\delta_2 \mid \phi)\ V_{\delta_1}(h_1)\ V_{\delta_2}(h_2), \end{aligned} \tag{3.3}$$

où $P(\phi = 0) = 1 - f$ et $P(\phi = 1) = f$. Les probabilités conditionnelles $P(T = \delta_1\delta_2 \mid \phi)$ sont aisément déduites du tableau 3.1 :

$$P(T=00 \mid \phi=0) = \frac{(1-f_0)\,(1-p)^2}{1-f}; \tag{3.4.a}$$

$$P(T=01 \mid \phi=0) = \frac{(1-f_1)\,p(1-p)}{1-f}; \tag{3.4.b}$$

$$P(T=10 \mid \phi=0) = \frac{(1-f_1)\,p(1-p)}{1-f}; \tag{3.4.c}$$

$$P(T=11 \mid \phi=0) = \frac{(1-f_2)\,p^2}{1-f}; \tag{3.4.d}$$

$$P(T=00 \mid \phi=1) = \frac{f_0\,(1-p)^2}{f}; \tag{3.4.e}$$

$$P(T=01 \mid \phi=1) = \frac{f_1\,p(1-p)}{f}; \tag{3.4.f}$$

$$P(T=10 \mid \phi=1) = \frac{f_1\,p(1-p)}{f}; \tag{3.4.g}$$

$$P(T=11 \mid \phi=1) = \frac{f_2\,p^2}{f}. \tag{3.4.h}$$

Ainsi, la probabilité conjointe du diplotype $d = [h_1,h_2]$ d'un individu et de son phénotype est déduite en sommant sur ses 4 diplotypes possibles au TIM :

$$\begin{aligned} P(d,\, \phi \mid V_0,\, V_1) \;&=\; \sum_T P(T = \delta_1\delta_2,\; \phi \mid [h_1,h_2],\, V_0,\, V_1) \\ &=\; P(\phi) \sum_T P(T = \delta_1\delta_2 \mid \phi)\; V_{\delta_1}(h_1)\; V_{\delta_2}(h_2) \qquad (3.5) \\ &=\; P(\phi) \Big[P(T = 00 \mid \phi)\; V_0(h_1)\; V_0(h_2) \\ &\qquad +\quad P(T = 01 \mid \phi)\; V_0(h_1)\; V_1(h_2) \\ &\qquad +\quad P(T = 10 \mid \phi)\; V_1(h_1)\; V_0(h_2) \\ &\qquad +\quad P(T = 11 \mid \phi)\; V_1(h_1)\; V_1(h_2) \Big]. \end{aligned}$$

Finalement, on obtient la probabilité d'un diplotype T au TIM, conditionnelle au phénotype et au diplotype $d = [h_1,h_2]$, par le quotient des équations 3.3 et 3.5 :

$$P(T = \delta_1\delta_2 \mid d,\, \phi,\, V_0,\, V_1) \;=\; \frac{P(T = \delta_1\delta_2 \mid \phi)\; V_{\delta_1}(h_1)\; V_{\delta_2}(h_2)}{\sum_T P(T = \delta_1\delta_2 \mid \phi)\; V_{\delta_1}(h_1)\; V_{\delta_2}(h_2)}. \tag{3.6}$$

Notons maintenant $n_{d,\phi}$ le nombre d'individus dans notre échantillon portant le diplotype d et le phénotype ϕ. L'étape E de l'algorithme EM consiste donc à estimer, pour tout h^δ,

$$\begin{aligned} m_{h^\delta}^{(k+1)} \;=\; & \sum_{(d=[h,h_2],\,\phi)\,\in\,(D,\Phi)} \left(n_{d,\phi} \sum_{\delta_2\in\{0,1\}} P\left(T = \delta\delta_2 \,\middle|\, d,\, \phi,\, V_0^{(k)},\, V_1^{(k)}\right)\right) \\ & + \sum_{(d=[h_1,h],\,\phi)\,\in\,(D,\Phi)} \left(n_{d,\phi} \sum_{\delta_1\in\{0,1\}} P\left(T = \delta_1\delta \,\middle|\, d,\, \phi,\, V_0^{(k)},\, V_1^{(k)}\right)\right). \end{aligned} \tag{3.7}$$

3.2.3 Échantillon stratifié

La méthode d'estimation des allèles au TIM décrite ci-dessus suppose un échantillon aléatoire simple d'individus diploïdes tirés de la population. Cependant les données portant sur un phénotype rare se présentent rarement sous cette forme, car cela nécessiterait un échantillon de très grande taille afin d'obtenir assez de cas. Un échantillon typique consiste plutôt en un certain nombre de cas et de témoins, respectivement tirés parmi les cas et les témoins de la population, et où $n_{\text{cas}} \approx n_{\text{témoins}}$. Par conséquent, le modèle de pénétrance F et la distribution des allèles au TIM dans la population (tableau 3.1) ne reflètent pas ce que l'on s'attend à observer dans un échantillon ainsi stratifié. Néanmoins, Boucher a théoriquement démontré que la présente méthode était robuste à un tel échantillonnage.

3.2.4 Algorithme EM

Afin de bien saisir la structure de la méthode d'estimation des allèles au TIM que nous venons de présenter, en voici les grandes étapes :

I. Calculer les probabilités $P(T|\phi)$ (équations 3.4.a à 3.4.h) ;

II. Déterminer les distributions initiales $V_0^{(0)}$ et $V_1^{(0)}$ (si aucune information *a priori* n'est disponible, une distribution uniforme est préférable) ;

III. À chaque itération k, tant que le seuil de convergence n'est pas atteint :

Étape E

i. Pour chacun des diplotypes $(d,\, \phi) \in (D,\, \Phi)$:

Pour chacun des 4 diplotypes possibles au TIM ($T \in \{00,\ 01,\ 10,\ 11\}$) :

Calculer $P\left(T = \delta_1\delta_2 \mid d,\, \phi,\, V_0^{(k)},\, V_1^{(k)}\right)$ (équation 3.6) ;

ii. Pour tout haplotype h^δ :

Calculer $m_{h^\delta}^{(k+1)}$ (équation 3.7) ;

Étape M

Pour tout h :

Mettre à jour les distributions $V_0(h)^{(k+1)}$ et $V_1(h)^{(k+1)}$) (équations 3.2) ;

Test de convergence

Si $\max_{h^\delta} \left(\left| V_\delta(h)^{(k+1)} - V_\delta(h)^{(k)} \right| \right) < \epsilon = \frac{0{,}01}{2^L}$, terminer l'algorithme

(2^L est le nombre d'haplotypes possibles sur une séquence de L SNPs) ;

IV. $\hat{V}_0 = V_0(h)^{(k+1)}$ et $\hat{V}_1 = V_1(h)^{(k+1)}$.

3.2.5 Exemple simple

Nous allons illustrer l'algorithme à l'aide d'un exemple simple. Soit un échantillon stratifié de 20 individus diploïdes témoins ($\phi = 0$) et 20 cas ($\phi = 1$), génotypés sur 2 SNPs α et β. Nous noterons $\{A,a\}$ les allèles possibles au SNP α et $\{B,b\}$ ceux au SNP β. Nous supposerons un modèle de pénétrance $F = \{0{,}01,\ 0{,}05,\ 0{,}1\}$, une fréquence dans la population du phénotype *cas* de $f = 0{,}0181$ et une fréquence de l'allèle $\delta = 1$ au TIM de $p = 0{,}1$. Le tableau 3.2 montre la distribution des diplotypes de cet échantillon.

Tableau 3.2 Décompte des diplotypes de notre échantillon.

$d = [h_1, h_2]$	$\phi = 0$	$\phi = 1$
$[\ ab\ ,\ ab\ ]$	3	1
$[\ ab\ ,\ aB\ ]$	3	1
$[\ aB\ ,\ aB\ ]$	2	2
$[\ ab\ ,\ Ab\ ]$	3	1
$[\ ab\ ,\ AB\ ]$	2	2
$[\ Ab\ ,\ aB\ ]$	2	2
$[\ aB\ ,\ AB\ ]$	1	3
$[\ Ab\ ,\ Ab\ ]$	2	2
$[\ Ab\ ,\ AB\ ]$	1	3
$[\ AB\ ,\ AB\ ]$	1	3
Total	20	20

I. Calculer les probabilités $P(T|\phi)$ (équations 3.4.a à 3.4.h) :

$$
\begin{aligned}
P(T=00 \mid \phi=0) &= \frac{(1-f_0)\,(1-p)^2}{1-f} = \frac{(1-0.01)\,(1-0.1)^2}{1-0.0181} = 0.8167;\\
P(T=01 \mid \phi=0) &= \frac{(1-f_1)\,p(1-p)}{1-f} = \frac{(1-0.05)\,0.1(1-0.1)}{1-0.0181} = 0.0871;\\
P(T=10 \mid \phi=0) &= \frac{(1-f_1)\,p(1-p)}{1-f} = \frac{(1-0.05)\,0.1(1-0.1)}{1-0.0181} = 0.0871;\\
P(T=11 \mid \phi=0) &= \frac{(1-f_2)\,p^2}{1-f} = \frac{(1-0.1)\,0.1^2}{1-0.0181} = 0.0092;\\
P(T=00 \mid \phi=1) &= \frac{f_0\,(1-p)^2}{f} = \frac{0.01\,(1-0.1)^2}{0.0181} = 0.4475;\\
P(T=01 \mid \phi=1) &= \frac{f_1\,p(1-p)}{f} = \frac{0.05\,(0.1)(1-0.1)}{0.0181} = 0.2486;\\
P(T=10 \mid \phi=1) &= \frac{f_1\,p(1-p)}{f} = \frac{0.05\,(0.1)(1-0.1)}{0.0181} = 0.2486;\\
P(T=11 \mid \phi=1) &= \frac{f_2\,p^2}{f} = \frac{0.1\,(0.1)^2}{0.0181} = 0.0552.
\end{aligned}
$$

II. Déterminer les distributions initiales $V_0^{(0)}$ et $V_1^{(0)}$:

Nous utiliserons les distributions uniformes présentées au tableau 3.3.

Tableau 3.3 Distributions initiales $V_0^{(0)}$ et $V_1^{(0)}$.

h	$V_0^{(0)}$	$V_1^{(0)}$
ab	0,25	0,25
aB	0,25	0,25
Ab	0,25	0,25
AB	0,25	0,25

III. À chaque itération k, tant que le seuil de convergence n'est pas atteint :

Étape E

i. Pour chacun des diplotypes $(d,\, \phi) \in (D,\, \Phi)$:

Pour chacun des 4 diplotypes possibles au TIM ($T \in \{00,\ 01,\ 10,\ 11\}$) :

Calculer $P\left(T = \delta_1\delta_2 \mid d,\, \phi,\, V_0^{(k)},\, V_1^{(k)}\right)$ (équation 3.6) :

Prenons le diplotype $(d = [ab,\, ab],\, \phi = 0)$. Calculons d'abord ses 4 $P(T = \delta_1\delta_2,\ \phi \mid d,\, V_0,\, V_1)$ (équation 3.3) :

$$
\begin{aligned}
P\left(T = 00, \phi = 0 \middle| d = [ab,ab], V_0^{(0)}, V_1^{(0)}\right) &= P(\phi = 0) P(T = 00 | \phi = 0) V_0^{(0)}(ab) V_0^{(0)}(ab) \\
&= 0.8167\ (0{,}25)\ (0{,}25)\ P(\phi = 0) \\
&= \mathbf{0.051}\ P(\phi = 0);
\end{aligned}
$$

$$
\begin{aligned}
P\left(T=01,\phi=0\middle|d=[ab,ab],V_0^{(0)},V_1^{(0)}\right) &= P(\phi=0)P(T=01|\phi=0)V_0^{(0)}(ab)V_1^{(0)}(ab)\\
&= 0.0871\ (0{,}25)\ (0{,}25)\ P(\phi=0)\\
&= \mathbf{0.0054}\ P(\phi=0);\\
P\left(T=10,\phi=0\middle|d=[ab,ab],V_0^{(0)},V_1^{(0)}\right) &= P(\phi=0)P(T=10|\phi=0)V_1^{(0)}(ab)V_0^{(0)}(ab)\\
&= 0.0871\ (0{,}25)\ (0{,}25)\ P(\phi=0)\\
&= \mathbf{0.0054}\ P(\phi=0);\\
P\left(T=11,\phi=0\middle|d=[ab,ab],V_0^{(0)},V_1^{(0)}\right) &= P(\phi=0)P(T=11|\phi=0)V_1^{(0)}(ab)V_1^{(0)}(ab)\\
&= 0.0092\ (0{,}25)\ (0{,}25)\ P(\phi=0)\\
&= \mathbf{0.0006}\ P(\phi=0),
\end{aligned}
$$

et enfin leur somme (équation 3.5) :

$$P\left(d=[ab,ab],\phi=0\middle|V_0^{(0)},V_1^{(0)}\right) = P(\phi=0)\sum_T P(T=\delta_1\delta_2|\phi=0)V_{\delta_1}^{(0)}(ab)V_{\delta_2}^{(0)}(ab)$$

$$= P(\phi=0)\,(0.051+0.0054+0.0054+0.0006)$$

$$= \mathbf{0.0625}\;P(\phi=0),$$

ce qui nous permet de calculer chacune de ses 4 $P(T=\delta_1\delta_2 \mid d,\,\phi,\,V_0,\,V_1)$ (équation 3.6) :

$$P\left(T=00 \mid d=[ab,\,ab],\,\phi=0,\,V_0^{(0)},\,V_1^{(0)}\right) = \frac{0.051}{0.0625} = \mathbf{0.8167};$$

$$P\left(T=01 \mid d=[ab,\,ab],\,\phi=0,\,V_0^{(0)},\,V_1^{(0)}\right) = \frac{0.0054}{0.0625} = \mathbf{0.0871};$$

$$P\left(T=10 \mid d=[ab,\,ab],\,\phi=0,\,V_0^{(0)},\,V_1^{(0)}\right) = \frac{0.0054}{0.0625} = \mathbf{0.0871};$$

$$P\left(T=11 \mid d=[ab,\,ab],\,\phi=0,\,V_0^{(0)},\,V_1^{(0)}\right) = \frac{0.0006}{0.0625} = \mathbf{0.0092}.$$

Les calculs sont ainsi effectués pour chacun des 19 autres diplotypes $(d,\, \phi) \in (D,\, \Phi)$ (tableau 3.2). Les probabilités que nous venons d'estimer nous permettent de calculer que, pour 3 individus $(d = [ab,\, ab],\, \phi = 0)$, il y en aura en moyenne $3 \times 0.8167 = 2.45$ qui porteront l'allèle $\delta = 0$ sur chacun de leurs deux haplotypes ab.

ii. Pour tout haplotype h^δ :

Calculer $m_{h^\delta}^{(k+1)}$ (équation 3.7) :

Par exemple, $m_{ab^0}^{(1)}$ est obtenu en sommant les nombres moyens d'haplotypes ab^0 retrouvés pour tous les diplotypes $(d,\, \phi) \in (D,\, \Phi)$.

Le tableau 3.4 donne le décompte moyen de tous les haplotypes h^δ ainsi obtenus après 1 itération.

Tableau 3.4 Décompte moyen des haplotypes après 1 itération.

h	$m_{h^0}^{(1)}$	$m_{h^1}^{(1)}$
ab	16,8294	3,1706
aB	15,9989	4,0011
Ab	15,9989	4,0011
AB	15,1684	4,8316
Total	63,9956	16,0044

Étape M

Pour tout h :

Mettre à jour les distributions $V_0(h)^{(k+1)}$ et $V_1(h)^{(k+1)}$ (équations 3.2) :

Par exemple, on aura que

$$V_0(ab)^{(1)} = \frac{m_{ab^0}^{(1)}}{m_0^{(1)}} = \frac{16{,}8294}{63{,}9956} = 0{,}2630.$$

Le tableau 3.5 donne les distributions estimées après 1 itération.

Test de convergence

Si $\max_{h^\delta} \left(\left| V_\delta(h)^{(k+1)} - V_\delta(h)^{(k)} \right| \right) < \epsilon$, terminer l'algorithme :

Ici, $\max_{h^\delta} \left(\left| V_\delta(h)^{(1)} - V_\delta(h)^{(0)} \right| \right) = 0{,}0519 > \epsilon = \frac{0{,}01}{4} = 0{,}0025$, donc on continue l'algorithme à l'étape E. Dans notre exemple, le plus grand incrément de convergence deviendra inférieur à notre seuil $\epsilon = 0{,}0025$ après la 27^e itération.

Tableau 3.5 Distributions estimées après 1 itération.

h	$\hat{V}_0^{(1)}$	$\hat{V}_1^{(1)}$
ab	0,2630	0,1981
aB	0,2500	0,2500
Ab	0,2500	0,2500
AB	0,2370	0,3019

IV. $\hat{V}_0 = V_0(h)^{(k+1)}$ et $\hat{V}_1 = V_1(h)^{(k+1)}$:

Le tableau 3.6 donne les distributions finales estimées après 27 itérations.

Tableau 3.6 Distributions finales estimées après 27 itérations.

h	$\hat{V}_0^{(27)}$	$\hat{V}_1^{(27)}$
ab	0,3150	0,0005
aB	0,2791	0,1382
Ab	0,2791	0,1382
AB	0,1268	0,7231

3.2.6 Implantation dans MapARG

Afin de bien saisir l'utilisation de cet algorithme EM par la méthode MapARG, voici où il s'insère dans l'algorithme de MapARG, ainsi que l'utilisation des distributions estimées $\hat{V}_0$ et $\hat{V}_1$ pour générer les allèles au TIM des haplotypes des individus d'un échantillon :

I. Choisir l'ensemble des positions x_T pour lesquelles $CL(x_T)$ sera évaluée ;

II. Pour chacune des $L - d + 1$ fenêtres couvrant l'ensemble des SNPs de l'échantillon :

   1. **Estimer les distributions V_0 et V_1 avec l'algorithme EM** (section 3.2.4) ;
   2. Pour chacun des $d - 1$ intervalles situés dans la fenêtre :

      Pour chacun des K graphes à construire :

      1. **Pour chaque individu de l'échantillon :**
         i. **Générer, avec $\hat{V}_0$, $\hat{V}_1$, les équations 3.4.a à 3.4.h, et par l'équation 3.6, la distribution de ses 4 T possibles ;**
         ii. **Générer $T = \delta_1\delta_2$ selon cette distribution ;**

2. Pour chaque étape τ du graphe, tant que le MRCA n'est pas atteint :
   i. Calculer $\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{P_{x_T}(H_{\tau+1}|H_\tau)} = \frac{\phi(H_\tau)}{\phi(H_{\tau+1})}$;
   ii. Mettre à jour $Q_{x_T}(H_\tau)$ et $P_{x_T}(H_{\tau+1})$;
   iii. Générer un évènement selon $P_{x_T}(H_{\tau+1})$;
   iv. Mettre à jour $H_{\tau+1}$;

III. Pour chaque position x_T :

Calculer $\hat{CL}(x_T)$;

IV. $\hat{x}_T$ correspond au maximum de $\hat{CL}(x_T)$.

3.3 Évaluation de la méthode

En présence d'un échantillon D, Φ composé d'individus pour lesquels on connait leur diplotype d et leur phénotype ϕ, mais pas leurs allèles au TIM ($T = \delta_1\delta_2$), l'efficacité de MapARG à bien estimer la position du TIM dépendra de l'exactitude de la méthode à estimer ces allèles.

3.3.1 Taux de succès

Afin d'évaluer l'efficacité de la méthode à estimer avec justesse l'allèle au TIM sur chacun des haplotypes, nous calculerons différents taux de succès. Définissons tout d'abord quelques termes. Soient :

n : le nombre total d'haplotypes dans l'échantillon (de n/2 individus) ;

n_{tem} : le nombre d'haplotypes provenant d'un *témoin* ($\phi = 0$) ;

n_{cas} : le nombre d'haplotypes provenant d'un *cas* ($\phi = 1$) ;

n^0 : le nombre d'haplotypes porteurs de l'allèle *primitif* au TIM ($\delta = 0$) ;

n^1 : le nombre d'haplotypes porteurs de l'allèle *mutant* au TIM ($\delta = 1$) ;

n^0_{tem} : le nombre d'haplotypes témoins porteurs de l'allèle primitif au TIM ;

n^1_{tem} : le nombre d'haplotypes témoins porteurs de l'allèle mutant au TIM ;

n^0_{cas} : le nombre d'haplotypes cas porteurs de l'allèle primitif au TIM ;

n^1_{cas} : le nombre d'haplotypes cas porteurs de l'allèle mutant au TIM.

Ainsi, nous avons : $n = n_{\text{tem}} + n_{\text{cas}} = n^0 + n^1 = n^0_{\text{tem}} + n^1_{\text{tem}} + n^0_{\text{cas}} + n^1_{\text{cas}}$.

Nous noterons aussi par $n_*^{*(0)}$ et $n_*^{*(1)}$ le nombre d'haplotypes qui seront estimés par la méthode comme étant porteurs, respectivement, de l'allèle 0 et 1 au TIM. Afin d'évaluer l'efficacité de la méthode à estimer correctement l'allèle au TIM, nous calculerons les *taux de succès partiels* :

$$\pi^0_{\mathsf{tem}} = \frac{n^{0\ (0)}_{\mathsf{tem}}}{n^0_{\mathsf{tem}}};$$

$$\pi^1_{\mathsf{tem}} = \frac{n^{1\ (1)}_{\mathsf{tem}}}{n^1_{\mathsf{tem}}};$$

$$\pi^0_{\mathsf{cas}} = \frac{n^{0\ (0)}_{\mathsf{cas}}}{n^0_{\mathsf{cas}}};$$

$$\pi^1_{\mathsf{cas}} = \frac{n^{1\ (1)}_{\mathsf{cas}}}{n^1_{\mathsf{cas}}},$$

les *taux de succès semi-partiels* :

$$\pi_{\mathsf{tem}} = \frac{n^0_{\mathsf{tem}}{}^{(0)} + n^1_{\mathsf{tem}}{}^{(1)}}{n^0_{\mathsf{tem}} + n^1_{\mathsf{tem}}} = \frac{n^0_{\mathsf{tem}}{}^{(0)} + n^1_{\mathsf{tem}}{}^{(1)}}{n_{\mathsf{tem}}};$$

$$\pi_{\mathsf{cas}} = \frac{n^0_{\mathsf{cas}}{}^{(0)} + n^1_{\mathsf{cas}}{}^{(1)}}{n^0_{\mathsf{cas}} + n^1_{\mathsf{cas}}} = \frac{n^0_{\mathsf{cas}}{}^{(0)} + n^1_{\mathsf{cas}}{}^{(1)}}{n_{\mathsf{cas}}};$$

$$\pi^0 = \frac{n^0_{\mathsf{tem}}{}^{(0)} + n^0_{\mathsf{cas}}{}^{(0)}}{n^0_{\mathsf{tem}} + n^0_{\mathsf{cas}}} = \frac{n^{0(0)}}{n^0};$$

$$\pi^1 = \frac{n^1_{\mathsf{tem}}{}^{(1)} + n^1_{\mathsf{cas}}{}^{(1)}}{n^1_{\mathsf{tem}} + n^1_{\mathsf{cas}}} = \frac{n^{1(1)}}{n^1},$$

analogues à la spécificité (π^0_{tem}, π^0_{cas} et π^0) et à la sensibilité (π^1_{tem}, π^1_{cas} et π^1) d'un test, en épidémiologie, ainsi que le *taux de succès global* :

$$\pi = \frac{n^0_{\mathsf{tem}}{}^{(0)} + n^1_{\mathsf{tem}}{}^{(1)} + n^0_{\mathsf{cas}}{}^{(0)} + n^1_{\mathsf{cas}}{}^{(1)}}{n^0_{\mathsf{tem}} + n^1_{\mathsf{tem}} + n^0_{\mathsf{cas}} + n^1_{\mathsf{cas}}} = \frac{n^{0(0)} + n^{1(1)}}{n}.$$

Cependant, ces taux de succès de la méthode à inférer l'allèle au TIM sur chacun des haplotypes ne reflètent pas nécessairement la proximité entre les distributions estimées et les distributions réelles des haplotypes porteurs des allèles 0 et 1 de l'échantillon. En effet, un allèle au TIM erronément inféré 1 sur un haplotype pourrait être compensé par un allèle erronément inféré 0 sur un autre haplotype identique. Pour cette raison, un taux de succès global *utilitaire* sera également calculé :

$$\pi_{\text{utilitaire}} \quad = \quad 1 \; - \; \frac{1}{n} \sum_h \left| {}^h n^0 - {}^h n^{(0)} \right| \quad = \quad 1 \; - \; \frac{1}{n} \sum_h \left| {}^h n^1 - {}^h n^{(1)} \right|,$$

où ${}^h n^\delta$ et ${}^h n^{(\delta)}$ sont les nombres d'haplotypes de type h qui sont, respectivement, *réellement* porteurs et *inférés* porteurs de l'allèle δ au TIM.

Pour bien saisir la subtilité, prenons un exemple simple d'un échantillon ne comprenant qu'un seul type h de 100 haplotypes, dont 50 sont porteurs de l'allèle 0 au TIM et 50 de l'allèle 1 :

	n^0	$n^{0(0)}$	$n^{0(1)}$	n^1	$n^{1(0)}$	$n^{1(1)}$	$n^{(0)}$	$n^{(1)}$
haplotype h	50	**40**	**10**	50	**20**	**30**	**60**	**40**

Puisque 70 haplotypes se voient correctement inférés leur allèle au TIM, le taux de succès global est donc égal à 0,7 :

$$\pi \quad = \quad \frac{n^{0(0)} + n^{1(1)}}{n} \quad = \quad \frac{\mathbf{40 + 30}}{100} \quad = \quad 0{,}7.$$

Par contre, on observe que 90 % de la distribution estimée résultante ($n^{(0)}$=60, $n^{(1)}$=40) correspond à la distribution réelle ($n^0 = 50$, $n^1 = 50$), et ce taux de succès global utilitaire est obtenu par :

$$
\begin{aligned}
\pi_{\text{utilitaire}} &= 1 - \frac{1}{n}\sum_h \left|{}^h n^0 - {}^h n^{(0)}\right| &&= 1 - \frac{1}{n}\sum_h \left|{}^h n^1 - {}^h n^{(1)}\right| \\
&= 1 - \frac{1}{100}\left|50 - 60\right| &&= 1 - \frac{1}{100}\sum_h \left|50 - 40\right| \\
&= 1 - 0{,}1 &&= 1 - 0{,}1 \\
&= 0{,}9.
\end{aligned}
$$

La figure 3.2 montre un exemple des courbes des 2 taux de succès globaux, des 4 taux partiels et des 4 taux semi-partiels obtenus sur un échantillon. L'algorithme EM est exécuté pour chacune des $L - d + 1$ fenêtres couvrant l'ensemble des SNPs de l'échantillon (voir son implantation dans MapARG, section 3.2.6). Ainsi, les 10 taux de succès sont calculés à chaque fenêtre, et positionnés sur les graphiques à mi-chemin entre les deux SNPs extrêmes de la fenêtre. La courbe d'un taux représente donc la proportion d'haplotypes dont l'allèle au TIM fut correctement inféré, pour chacune des fenêtres le long de la séquence.

Les 10 taux sont divisés en 4 graphiques afin de permettre une meilleure visualisation des résultats. La figure 3.2 montre les taux global π et global utilitaire $\pi_{\text{utilitaire}}$ (*a*), les 4 taux partiels π^0_{tem}, π^1_{tem}, π^0_{cas} et π^1_{cas} (*b*) ainsi que les taux semi-partiels π^0 et π^1 (*c*) et π_{tem} et π_{cas} (*d*).

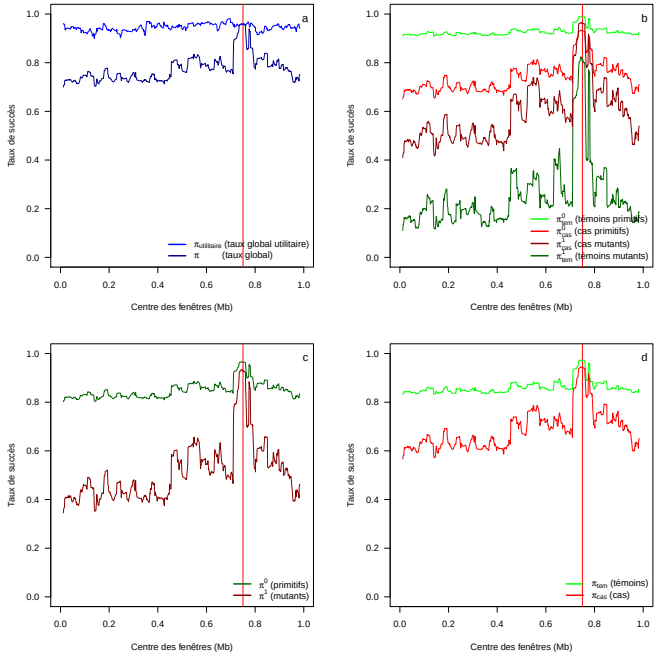


Figure 3.2 Exemple des taux de succès obtenus sur un échantillon. La droite verticale rouge indique la position du TIM recherché. *Détails techniques :* ces résultats ont été obtenus avec des fenêtres de 16 SNPs incrémentées de 1 sur 500 SNPs d'un échantillon de 400 cas et 400 témoins simulé selon $F = \{0{,}01,\ 0{,}1,\ 0{,}1\}$.

3.3.2 Facteurs testés

L'efficacité de la méthode à bien estimer les allèles au TIM sera testée en fonction de 4 facteurs, soient la largeur des fenêtres utilisées, la taille de l'échantillon et les *risques relatifs RR1* et *RR2* (tableau 3.7).

Tableau 3.7 Facteurs testés pour l'efficacité de la méthode.

Facteur	**Valeurs testées**			
RR1 ($\frac{f_1}{f_0}$)	1,01	1,1	2	10
RR2 ($\frac{f_2}{f_0}$)	1,01	1,1	2	10
Taille de l'échantillon ($\frac{n_{\text{tem}}}{2}$ / $\frac{n_{\text{cas}}}{2}$)	50/50	100/100	200/200	400/400
Largeur des fenêtres (SNPs)	2	4	8	16

Les risques relatifs RR1 et RR2 réfèrent à l'augmentation du risque d'un individu de développer le phénotype *cas* ($\phi = 1$) s'il possède 1 ($T = 01$ ou 10) ou 2 ($T = 11$) allèles mutants au TIM, respectivement, par rapport à s'il n'en possède aucun ($T = 00$). Ainsi,

$$\text{RR1} = \frac{f_1}{f_0} \qquad \text{et} \qquad \text{RR2} = \frac{f_2}{f_0}.$$

Plus les risques relatifs sont petits et se rapprochent de 1, plus l'information contenue dans les données et permettant de distinguer les haplotypes mutants (au TIM) des primitifs est subtile. Les 4 valeurs testées pour RR1 et RR2 seront 1,01, 1,1, 2 et 10. Afin de générer de tels échantillons, f_0 sera fixé à 0,01 alors que f_1 et f_2 prendront les valeurs 0,0101, 0,011, 0,02 et 0,1, valeurs réalistes en génétique. Étant logique d'assumer que $f_0 \leq f_1 \leq f_2$, seuls des échantillons respectant cette contrainte seront générés.

À titre indicatif, le tableau 3.8 donne les valeurs théoriques des nombres d'haplotypes que l'on devrait retrouver en moyenne dans un échantillon de 250/250 témoins/cas (1 000 haplotypes), en fonction de RR1 et RR2, obtenus par :

$$ {}^{*}n^{0}_{\text{tem}} = n_{\text{tem}} \frac{(1-p)^2(1-f_0) + p(1-p)(1-f_1)}{(1-p)^2(1-f_0) + 2p(1-p)(1-f_1) + p^2(1-f_2)}; $$

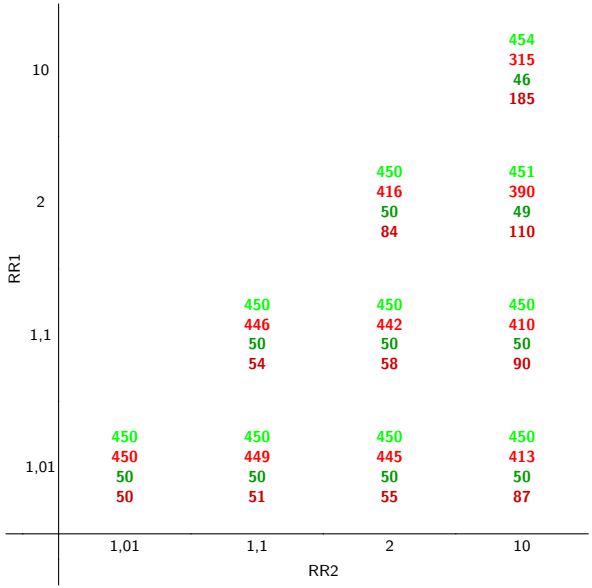
$$ {}^{*}n^{0}_{\text{cas}} = n_{\text{cas}} \frac{(1-p)^2 f_0 + p(1-p) f_1}{(1-p)^2 f_0 + 2p(1-p) f_1 + p^2 f_2}; $$

$$ {}^{*}n^{1}_{\text{tem}} = n_{\text{tem}} \frac{p(1-p)(1-f_1) + p^2(1-f_2)}{(1-p)^2(1-f_0) + 2p(1-p)(1-f_1) + p^2(1-f_2)}; $$

$$ {}^{*}n^{1}_{\text{cas}} = n_{\text{cas}} \frac{p(1-p) f_1 + p^2 f_2}{(1-p)^2 f_0 + 2p(1-p) f_1 + p^2 f_2}, $$

où $p = 0{,}1$ et $f_0 = 0{,}01$.

Tableau 3.8 Nombres théoriques d'haplotypes en fonction de RR1 et RR2, retrouvés dans un échantillon de 250/250 témoins/cas où $p = 0{,}1$ et $f_0 = 0{,}01$.
Vert : Témoins primitifs (${}^*n^0_{\text{tem}}$) ;
Rouge : Cas primitifs (${}^*n^0_{\text{cas}}$) ;
Vert foncé : Témoins mutants (${}^*n^1_{\text{tem}}$) ;
Rouge foncé : Cas mutants (${}^*n^1_{\text{cas}}$).



Comme il devient de plus en plus facile de génotyper plusieurs individus pour les mêmes marqueurs, il est pertinent de tester si la disponibilité d'échantillons de plus grandes tailles améliorera la performance de la méthode. La taille des échantillons, en $\frac{n}{2}$ individus, est donnée par $\frac{n_{\text{tem}}}{2}$ / $\frac{n_{\text{cas}}}{2}$. À partir d'une même population, des échantillons de 4 tailles différentes seront générés, soient : 50/50, 100/100, 200/200 et 400/400.

Puisque l'algorithme EM est exécuté pour chacune des fenêtres couvrant l'ensemble des SNPs de l'échantillon, la largeur des fenêtres (et donc la longueur des haplotypes partiels) en nombre de SNPs sera testée. On s'attend à ce que de trop petites fenêtres, et donc de trop courts haplotypes, ne permettent pas à l'algorithme de bien estimer les distributions. Le nombres de types d'haplotypes possibles augmentant exponentiellement avec le nombre de SNPs $\left(2^{\text{SNPs}}\right)$, on s'attend aussi à ce qu'avec des fenêtres trop larges, chaque type d'haplotype ne contienne que 0 ou 1 haplotype. Les fenêtres testées seront larges de 2, 4, 8 et 16 SNPs.

3.3.3 Résultats sur 1 population

Ainsi, à partir d'une même population de $50\,000$ haplotypes et de $10\,000$ SNPs, simulée à l'aide de *fastsimcoal* (Excoffier et Foll, 2011), 40 échantillons sont générés à l'aide de notre programme *Sample*, couvrant les 10 combinaisons possibles de RR1 et RR2, et pour chacune des 4 tailles. Pour tous ces 40 échantillons, les mêmes 500 SNPs sont choisis aléatoirement et conservés, et le même TIM, de fréquence $p \approx 0{,}1$, est utilisé pour créer les phénotypes selon les différents modèles de pénétrance. Suivant l'algorithme de MapARG (section 3.2.6), une valeur de $K = 20$ est utilisée.

La figure 3.3 présente les résultats de ces simulations pour les taux globaux π et $\pi_{\text{utilitaire}}$, par modèle de pénétrance (RR1 et RR2) pour un échantillon de 400 témoins et 400 cas, et à l'aide de fenêtres de 16 SNPs. La figure 3.4 présente les résultats de la même manière pour les taux partiels π^0_{tem}, π^1_{tem}, π^0_{cas} et π^1_{cas}. Les résultats homologues pour d'autres combinaisons de taille d'échantillon et de largeur de fenêtre peuvent être consultés aux appendices A et B.

De façon analogue, la figure 3.5 présente les résultats pour les taux globaux π et $\pi_{\text{utilitaire}}$, par taille d'échantillon et par largeur de fenêtre, pour le modèle de pénétrance $F = \{0{,}01,\ 0{,}1,\ 0{,}1\}$, c'est-à-dire pour lequel RR1 = RR2 = 10. La figure 3.6 présente ces résultats pour les taux partiels π^0_{tem}, π^1_{tem}, π^0_{cas} et π^1_{cas}. Les résultats homologues pour d'autres modèles de pénétrance peuvent être consultés à l'appendice C. De plus, les appendices D, E et F montrent de la même manière que les appendices A, B et C les résultats pour les taux semi-partiels π^0, π^1, π_{tem} et π_{cas}.

Les figures 3.7, 3.8 et 3.9 permettent de mieux visualiser l'effet d'un facteur à la fois en combinant dans un même graphique, pour chacun de ces 6 taux (2 taux par figure), les 4 valeurs des RRs combinés (*a,d*), de la taille de l'échantillon (*b,e*) et de la largeur des fenêtres (*c,f*). Pour une rangée donnée, la courbe bleue est la même, représentant le taux obtenu pour RR1 = RR2 = 10, 400/400 témoins/cas et des fenêtres de 16 SNPs.

Afin de faciliter l'analyse, des lignes pointillées sont ajoutées aux graphiques, correspondant aux taux de succès *aléatoires*, c'est-à-dire les taux de succès moyens qui devraient être obtenus aléatoirement en l'absence de tout effet génétique du TIM. Ils sont calculés de la façon suivante :

$$^{*}\pi \quad = \quad P(\hat{\delta} = \delta \,|\, \delta)\; P(\delta).$$

Ainsi, pour $p = 0{,}1$, on obtient les 10 taux aléatoires suivants :

$$
\begin{aligned}
{}^*\pi^0 = {}^*\pi^0_{\mathsf{tem}} = {}^*\pi^0_{\mathsf{cas}} &= P(\hat{\delta} = 0 \,|\, \delta = 0)\; P(\delta = 0) \\
&= (1-p)(1) \\
&= \mathbf{0{,}9};
\end{aligned}
$$

$$
\begin{aligned}
{}^*\pi^1 = {}^*\pi^1_{\mathsf{tem}} = {}^*\pi^1_{\mathsf{cas}} &= P(\hat{\delta} = 1 \,|\, \delta = 1)\; P(\delta = 1) \\
&= p(1) \\
&= \mathbf{0{,}1};
\end{aligned}
$$

$$
\begin{aligned}
{}^*\pi_{\mathsf{utilitaire}} = {}^*\pi = {}^*\pi_{\mathsf{tem}} = {}^*\pi_{\mathsf{cas}} &= P(\hat{\delta} = 0 \,|\, 0)\; P(0) \;+\; P(\hat{\delta} = 1 \,|\, \delta = 1)\; P(\delta = 1) \\
&= (1-p)(1-p) \;+\; (p)(p) \\
&= 0{,}81 \;+\; 0{,}01 \\
&= \mathbf{0{,}82}.
\end{aligned}
$$

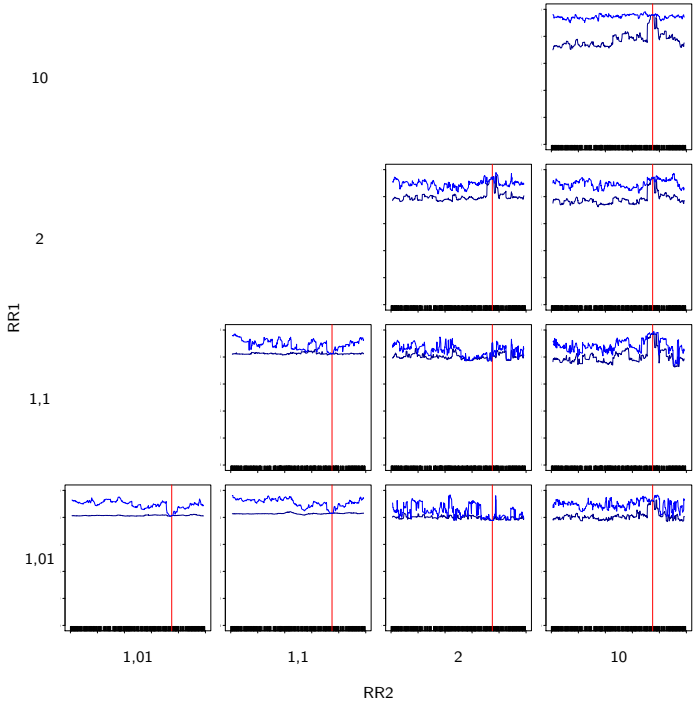


Figure 3.3 Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < > ≫

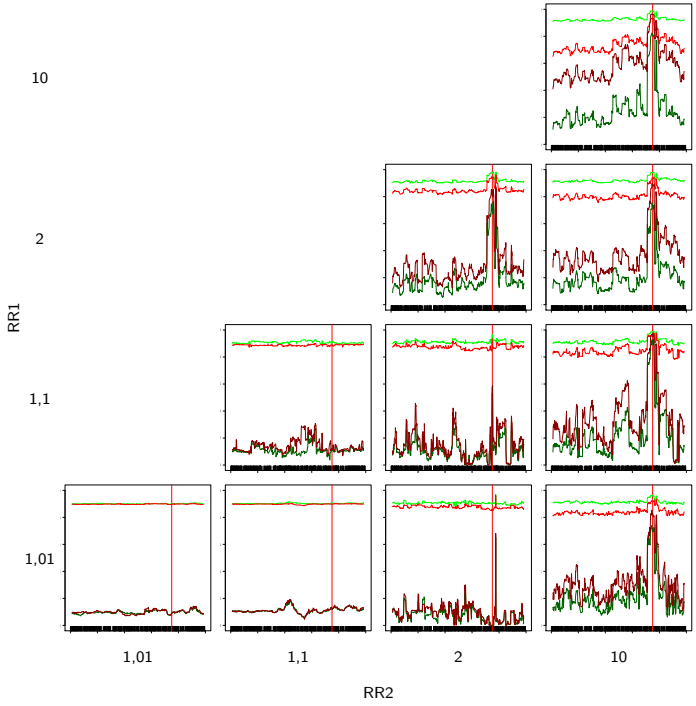


Figure 3.4 Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

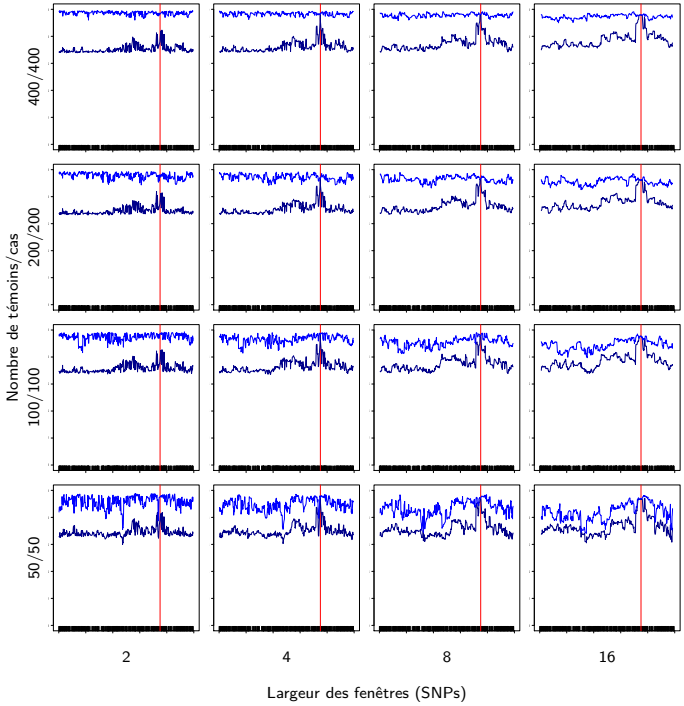


Figure 3.5 Taux de succès globaux en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = 10 et RR2 = 10. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < > ≫

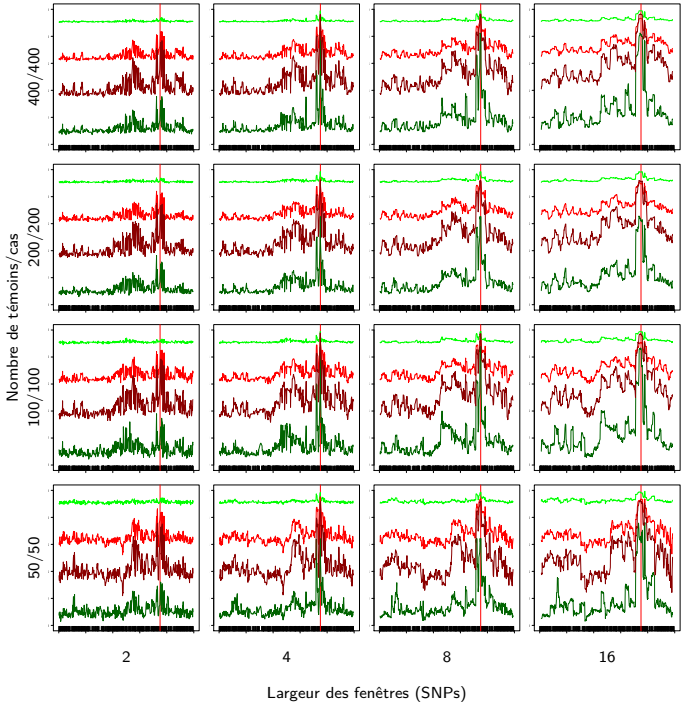


Figure 3.6 Taux de succès partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = 10 et RR2 = 10. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

≪ < > ≫

Tout d'abord, les figures 3.3 et 3.5 montrent des taux de succès globaux passablement élevés ($\pi \approx$ 0,7-0,9 ; $\pi_{\text{utilitaire}} \approx$ 0,8-1,0). Une analyse plus détaillée nous permet de constater l'effet de chacun des facteurs (figure 3.7). Pour ce qui est du taux $\pi_{\text{utilitaire}}$, une combinaison de très forts RRs (10) résulte en un meilleur taux que pour des RRs plus faibles (*a*). De même, le taux $\pi_{\text{utilitaire}}$ augmente graduellement avec la taille de l'échantillon (*b*). Cependant, la largeur des fenêtres semble avoir l'effet contraire, le taux augmentant graduellement avec des fenêtres de moins en moins larges (*c*). On remarque que le taux $\pi_{\text{utilitaire}}$ est quasiment toujours au-dessus du taux aléatoire ($^{*}\pi_{\text{utilitaire}} = 0{,}82$), sauf avec le plus petit échantillon de 50/50 témoins/cas.

Étrangement, le taux global π n'est pas affecté de la même façon que le taux $\pi_{\text{utilitaire}}$ par les différents facteurs. Alors que le taux global π montre clairement un pic dans la région du TIM, celui-ci disparait dans les distributions résultantes ($\pi_{\text{utilitaire}}$). Les risques relatifs (*d*) ont un effet net sur le taux global π. Alors que le taux est relativement plat et très

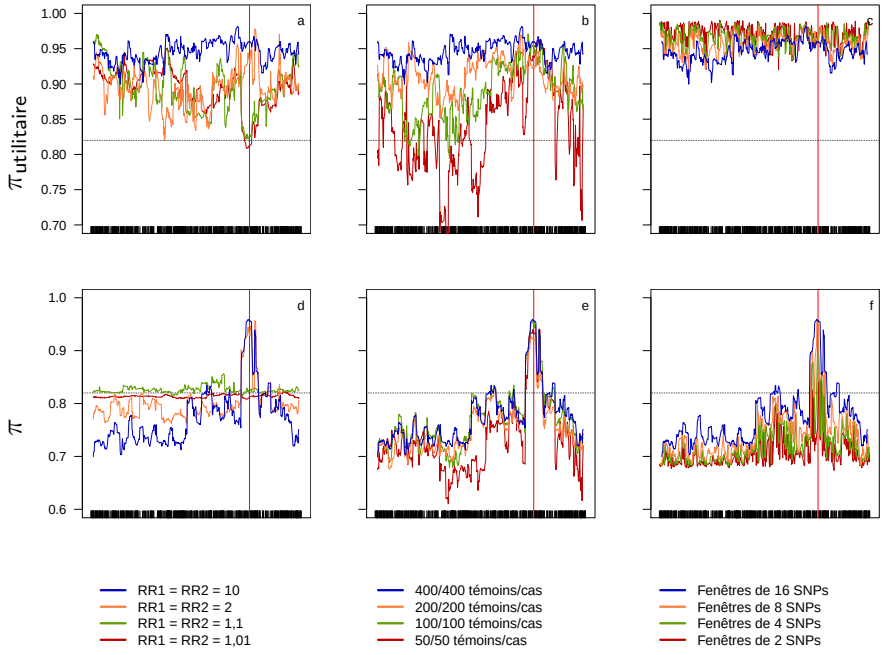


Figure 3.7 Taux de succès global utilitaire (*a*,*b*,*c*) et global (*d*,*e*,*f*) en fonction des risques relatifs RR1 et RR2 combinés (*a*,*d*), de la taille de l'échantillon (*b*,*e*) et de la largeur des fenêtres (*c*,*f*). Pour une rangée donnée, l'échelle des ordonnées est la même. La ligne pointillée représente le taux de succès aléatoire (0,82).
a,*d* : 400/400 témoins/cas, fenêtres de 16 SNPs ; *b*,*e* : RR1 = RR2 = 10, fenêtres de 16 SNPs ; *c*,*f* : RR1 = RR2 = 10, 400/400 témoins/cas.

près du taux aléatoire (${}^*\pi = 0{,}82$) pour des RR faibles (1,01 et 1,1), un pic, d'envergure équivalente, apparait au TIM pour des RRs forts (2 et 10), et le taux diminue en dessous du taux aléatoire sur le reste de la séquence, proportionnellement aux RRs. La taille de l'échantillon (*e*) ne semble avoir aucun effet sur π, à l'exception d'une légère baisse pour un très petit échantillon (50/50 témoins/cas). Enfin, le taux global π s'améliore graduellement avec la largeur des fenêtres utilisées (*f*), exactement le contraire du taux $\pi_{\text{utilitaire}}$.

Alors que le taux de succès global π est relativement élevé, sa décortication en ses 4 taux de succès partiels nous montre une inégalité flagrante. Les figures 3.4 et 3.6 nous permettent de visualiser leur grande divergence. Pour des RRs très faibles (figure 3.4), les 4 taux sont relativement plats le long de la séquence et suivent de très près leurs taux aléatoires respectifs et divergents (primitifs : 0,9 ; mutants : 0,1). À mesure que les RRs augmentent, les taux de succès des haplotypes mutants (π^1_{tem}, *vert foncé* ; π^1_{cas}, *rouge foncé*) s'améliorent considérablement alors que celui des cas primitifs (π^0_{cas}, *rouge*) se détériore. La

figure 3.6 semble montrer un effet plus subtile de la taille de l'échantillon et de la largeur des fenêtres sur ces taux de succès partiels (pour RR1 = RR2 = 10).

La figure 3.8 permet de comparer le taux de succès des haplotypes témoins primitifs (π^0_{tem}) et mutants (π^1_{tem}). Alors que l'amplitude de ces deux taux diverge considérablement, leurs courbes le long de la séquence sont très similaires, et l'effet des différents facteurs est très semblable. Comme pour le taux global π, le pic autour du TIM n'apparait que pour des RRs élevés (*a*,*d*). Par contre, les taux s'améliorent aussi sur le reste de la séquence, au-dessus de leurs taux aléatoires respectifs, contrairement au taux global π. Comme pour π cependant, la taille de l'échantillon ne semble exercer aucune influence sur les taux de succès des témoins (*b*,*e*), et la largeur des fenêtres exerce un effet positif graduel (*c*,*f*). On remarque que ces deux taux ne se trouvent quasiment jamais en dessous de leurs taux aléatoires respectifs.

De la même façon que pour les témoins, la figure 3.9 permet de comparer le taux de succès

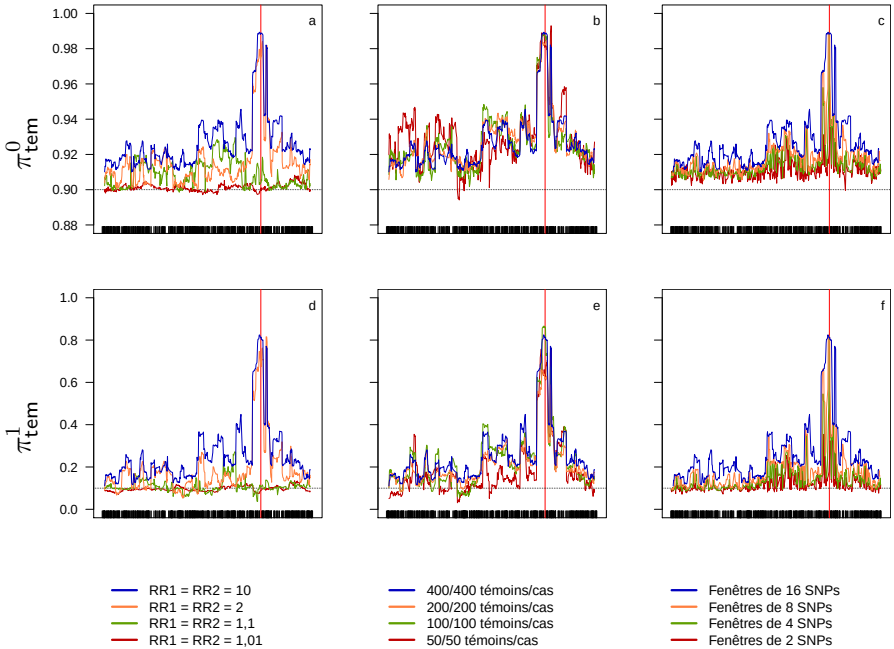


Figure 3.8 Taux de succès des témoins primitifs (*a*,*b*,*c*) et des témoins mutants (*d*,*e*,*f*) en fonction des risques relatifs RR1 et RR2 combinés (*a*,*d*), de la taille de l'échantillon (*b*,*e*) et de la largeur des fenêtres (*c*,*f*). Pour une rangée donnée, l'échelle des ordonnées est la même. La ligne pointillée représente le taux de succès aléatoire (*a*,*b*,*c* : 0,9 ; *d*,*e*,*f* : 0,1).
a,*d* : 400/400 témoins/cas, fenêtres de 16 SNPs ; *b*,*e* : RR1 = RR2 = 10, fenêtres de 16 SNPs ; *c*,*f* : RR1 = RR2 = 10, 400/400 témoins/cas.

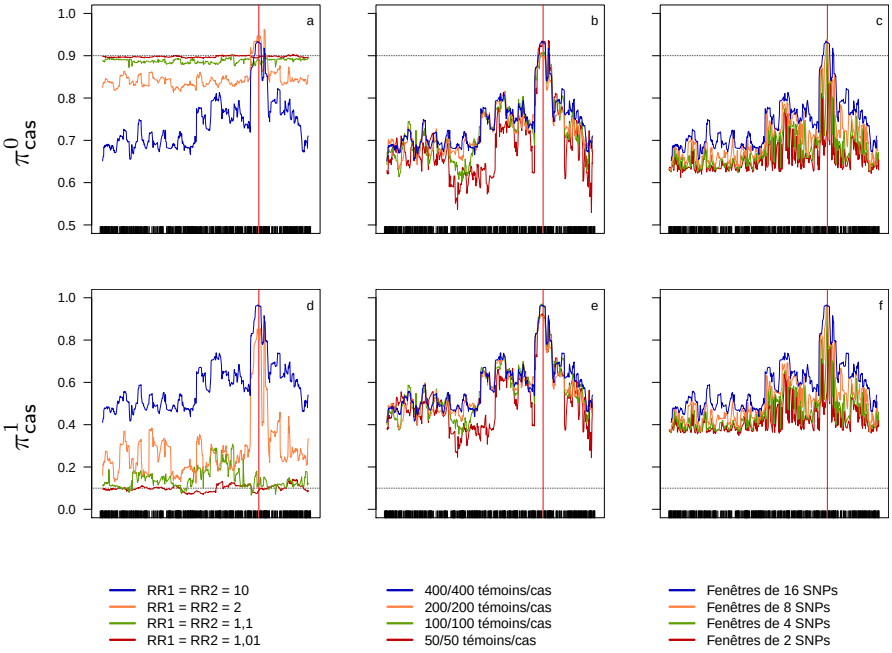


Figure 3.9 Taux de succès des cas primitifs (*a*,*b*,*c*) et des cas mutants (*d*,*e*,*f*) en fonction des risques relatifs RR1 et RR2 combinés (*a*,*d*), de la taille de l'échantillon (*b*,*e*) et de la largeur des fenêtres (*c*,*f*). Pour une rangée donnée, l'échelle des ordonnées est la même. La ligne pointillée représente le taux de succès aléatoire (*a*,*b*,*c* : 0,9 ; *d*,*e*,*f* : 0,1).
a,*d* : 400/400 témoins/cas, fenêtres de 16 SNPs ; *b*,*e* : RR1 = RR2 = 10, fenêtres de 16 SNPs ; *c*,*f* : RR1 = RR2 = 10, 400/400 témoins/cas.

des haplotypes cas primitifs (π^0_{cas}) et mutants (π^1_{cas}). Comme pour les témoins, la similitude des courbes de ces deux taux est frappante, mais dans des amplitudes différentes, pour les différentes tailles d'échantillon (*b,e*) et les différentes largeurs de fenêtre (*c,f*). De plus, leur effet est également très similaire à celui exercé sur le taux global π. Par contre, l'effet des risques relatifs diverge (*a,d*). Alors que le taux des cas mutants (π^1_{cas}) se comporte similairement à celui des témoins (π^0_{tem} et π^1_{tem}) mais de façon plus prononcée, seul le taux des cas primitifs (π^0_{cas}) est affecté de façon similaire au taux global π, se détériorant avec des RRs plus forts, à l'exception du pic qui se forme au TIM. On remarque d'ailleurs que ce taux π^0_{cas} se trouve toujours sous son taux aléatoire (0,9), à l'exception de son pic au TIM.

3.3.4 Résultats sur 100 populations

Afin de réaliser une analyse plus robuste de l'effet de nos 4 facteurs sur la performance de la méthode d'estimation des allèles au TIM, les mêmes simulations que celles exécutées à la

section 3.3.3 (incluant la génération de 40 échantillons) sont répétées sur 100 populations différentes. Dans le but de faciliter la visualisation de ces résultats, une mesure sera calculée pour chacune des 10 courbes de taux succès. Chacune de ces 10 courbes représente un taux de succès calculé pour une fenêtre qui est glissée le long de la séquence. Puisque les fenêtres sont glissées par incréments de 1 SNP, une fenêtre de 8 SNPs englobera 7 fois un intervalle donné entre 2 SNPs consécutifs. Le TIM étant positionné dans un tel intervalle, il sera donc englobé 7 fois par une telle fenêtre de 8 SNPs. Un taux de succès *périTIM* consistera donc en la moyenne de ce taux, calculée sur ces (ici 7) fenêtres qui englobent ainsi le TIM. Ces taux de succès périTIMs sont d'autant plus pertinents qu'ils ciblent les fenêtres les plus importantes pour le bon fonctionnement de la méthode MapARG à localiser le TIM. La moyenne de ces taux périTIMs sera ensuite calculée sur les 100 populations.

La figure 3.10 montre les résultats obtenus pour les 2 taux globaux et les 4 taux partiels en fonction des risques relatifs combinés (RR1 = RR2). On y voit l'évolution des taux

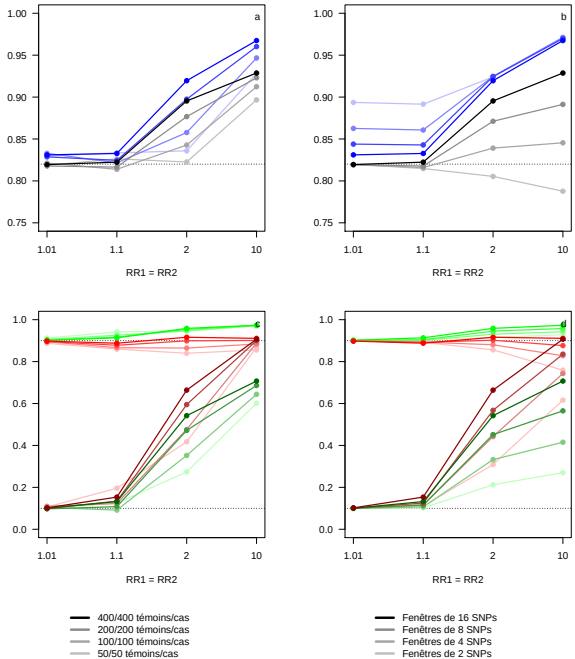


Figure 3.10 Taux de succès périTIMs en fonction des risques relatifs RR1 et RR2 combinés, par taille de l'échantillon (a,c) et par largeur de fenêtre (b,d). Chaque point représente la moyenne d'un taux de succès périTIM sur les 100 populations. Les lignes pointillées représentent les taux de succès aléatoires (a,b : 0,82 ; c,d : 0,1 et 0,9). a,c : fenêtres de 16 SNPs ; b,d : 400/400 témoins/cas. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Noir :* Taux global (π) ; *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

périTIMs en fonction de la taille de l'échantillon, pour des fenêtres de 16 SNPs (*a,c*), ainsi que celle des taux périTIMs en fonction de la largeur des fenêtres, pour des échantillons de 400 témoins et 400 cas (*b,d*).

On remarque que pour des RRs faibles, les 4 taux de succès périTIMs partiels (*c,d*), ainsi que leur taux global π résultant (*a,b, noir*), ne semblent pas diverger le moindrement de leurs taux aléatoires respectifs (0,1, 0,9 et 0,82), peu importe la taille de l'échantillon (*a,c*, pour des fenêtres de 16 SNPs) ou la largeur des fenêtres (*b,d*, pour un échantillon de 400/400 témoins/cas). Trois des 4 taux partiels s'améliorent considérablement (π^1_{tem}, *vert foncé* ; π^1_{cas}, *rouge foncé*) ou légèrement (π^0_{tem}, *vert*) avec des RRs forts (2 et 10). L'amélioration de π^1_{tem} et π^1_{cas} est proportionnelle à la taille de l'échantillon (*c*) pour RRs $= 2$, mais cet effet de la taille de l'échantillon s'estompe avec de très forts RRs (10) et semble inexistante pour π^0_{tem}. La largeur des fenêtres exerce quant à elle un effet graduel positif sur ces 3 taux partiels, qui persiste pour RRs $= 10$ (*d*).

Encore une fois, le taux de succès des cas primitifs π^0_{cas} (*rouge*) se comporte de façon inattendue. Quoiqu'insensible aux RRs pour des échantillons suffisamment grands et des fenêtres suffisamment larges, son taux périTIM se détériore avec l'augmentation des RRs avec de petits échantillons (*c*) ou dans une plus large mesure avec de courtes fenêtres (*d*).

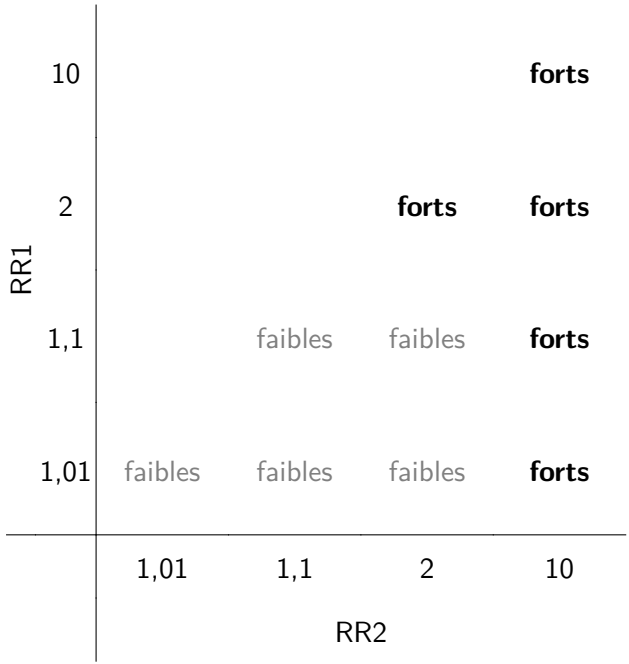
Le comportement du taux de succès global périTIM π (*noir*) reflète celui des 4 taux partiels qui le constituent, se délogeant de son taux aléatoire si les RRs sont suffisamment forts. Il s'améliore proportionnellement à la taille des échantillons (*a*) pour RRs $= 2$ (sauf pour un petit échantillon de 50/50), et cet effet de la taille de l'échantillon s'estompe avec de très forts RRs (10). L'effet graduel de la largeur des fenêtres (*b*) est quant à lui persistant pour des RRs de 10. Cependant, la dégradation du taux partiel des cas primitifs (π^0_{cas}) avec de courtes fenêtres semble assez significative pour affecter négativement le taux global π, qui se détériore avec l'augmentation des RRs pour de très courtes fenêtres de 2 SNPs (*b*).

Le taux de succès global périTIM $\pi_{\text{utilitaire}}$ (*bleu*) montre une configuration plutôt étrange,

la taille des échantillons et la largeur des fenêtres n'exerçant pas la même influence sur lui. Pour des fenêtres de 16 SNPs (*a*), le taux $\pi_{\text{utilitaire}}$ suit la même configuration que le taux π. Légèrement supérieur à son taux aléatoire pour des RRs faibles, peu importe la taille de l'échantillon, il s'améliore considérablement pour de forts RRs, proportionnellement à la taille de l'échantillon. Cependant, pour un échantillon de 400/400 témoins/cas (*b*), il est meilleur avec de courtes fenêtres pour des RRs faibles, et cette divergence disparait totalement avec des RRs forts, où la largeur des fenêtres n'exerce plus aucun effet. En somme, pour de forts RRs (2 et 10), il semble que de grands échantillons soient bénéfiques au taux $\pi_{\text{utilitaire}}$, alors que la largeur des fenêtres n'a pas d'importance, et que pour de faibles RRs (1,01 et 1,1), la taille des échantillons n'a pas d'influence sur le taux $\pi_{\text{utilitaire}}$, alors que de très courtes fenêtres (2 SNPs) lui semblent bénéfiques.

Les figures 3.11, 3.12 et 3.13 permettent de visualiser sur les 2 taux globaux et sur les 4 taux partiels l'effet de la taille des fenêtres et de la largeur des fenêtres pour les 10 modèles de

Tableau 3.9 Segmentation arbitraire des modèles de pénétrance testés en fonction de la force des risques relatifs résultants.



pénétrance testés, selon que les RRs résultants sont considérés faibles ou forts (tableau 3.9).

On constate immédiatement une nette discrimination dans l'effet de ces facteurs, selon que les RRs sont faibles ou forts. La figure 3.11 nous confirme que le taux global périTIM π est amélioré par de grands échantillons (*c*, avec des fenêtres de 16 SNPs) ou par de larges fenêtres (*d*, avec des échantillons de 400/400) si les RRs sont forts (*lignes foncées*), mais que ces deux facteurs n'ont aucun effet lorsque les RRs sont faibles (*lignes pâles*), où le taux global π ne déroge alors presque pas de son taux aléatoire. Ce même effet de la taille de l'échantillon est observé sur le taux $\pi_{\text{utilitaire}}$ (*a*), mais la largeur des fenêtres exerce une influence très différente (*b*). Pour des RRs forts, le taux périTIM $\pi_{\text{utilitaire}}$ est élevé et supérieur au taux aléatoire, peu importe la largeur des fenêtres, alors qu'en présence d'échantillons aux RRs faibles, des fenêtres moins larges résultent en un meilleur taux $\pi_{\text{utilitaire}}$, alors que de plus larges fenêtres tendent à le rapprocher du taux aléatoire.

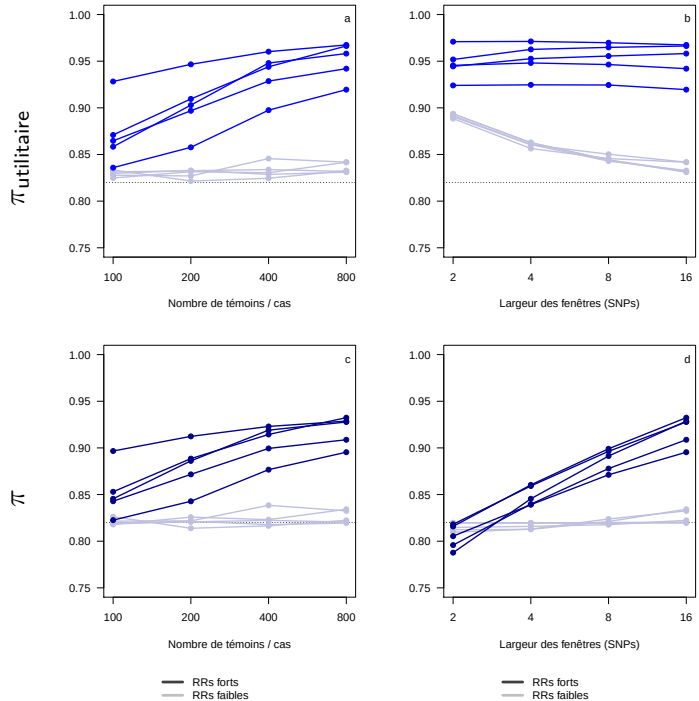


Figure 3.11 Taux de succès périTIMs globaux en fonction de la taille de l'échantillon (*a*,*c*) et de la largeur des fenêtres (*b*,*d*), par risques relatifs. Chaque point représente la moyenne d'un taux de succès périTIM sur les 100 populations. La ligne pointillée représente le taux de succès aléatoire (0,82). *a*,*c* : fenêtres de 16 SNPs ; *b*,*d* : 400/400 témoins/cas.

Les figures 3.12 et 3.13 permettent de constater que les taux partiels se comportent à peu près comme leur taux global résultant π, à quelques exceptions intéressantes près.

Avec de forts RRs (*lignes foncées*), alors que les 3 taux π^0_{tem}, π^1_{tem} et π^1_{cas} sont déjà supérieurs à leurs taux aléatoires respectifs pour de petits échantillons (50/50 témoins/cas) ou avec de petites fenêtres, seul le taux périTIM des cas primitifs (π^0_{cas}) s'y trouve inférieur, l'augmentation de l'un de ces deux facteurs le faisant passer au-dessus de son taux aléatoire. Aussi, on observe que pour certains modèles de pénétrance aux RRs faibles (*lignes pâles*), ces 3 mêmes taux π^0_{tem}, π^1_{tem} et π^1_{cas} sont légèrement améliorés par de larges fenêtres, avec des échantillons de 400/400 témoins/cas. Enfin, le taux π^0_{tem}, amélioré par de grands échantillons lorsque les RRs sont forts, se voit détérioré quand les RRs sont faibles.

Il est intéressant de noter que la plupart des courbes qui montrent une amélioration des taux en fonction de l'un des facteurs semblent tendre vers l'atteinte éventuelle d'un plateau.

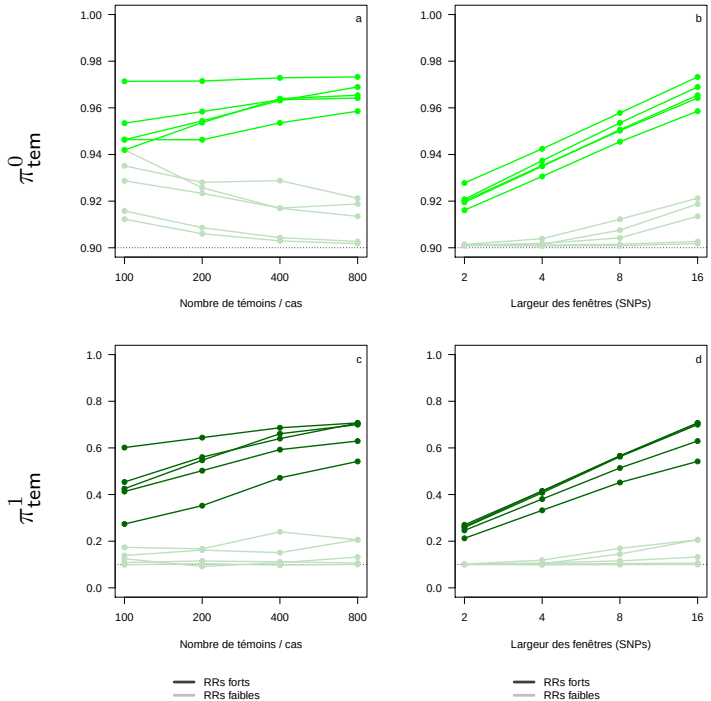


Figure 3.12 Taux de succès périTIMs des témoins primitifs (*a*,*b*) et mutants (*c*,*d*) en fonction de la taille de l'échantillon (*a*,*c*) et de la largeur des fenêtres (*b*,*d*), par risques relatifs. Chaque point représente la moyenne d'un taux de succès périTIM sur les 100 populations. La ligne pointillée représente le taux de succès aléatoire (*a*,*b* : 0,9 ; *c*,*d* : 0,1).
a,*c* : fenêtres de 16 SNPs ; *b*,*d* : 400/400 témoins/cas.

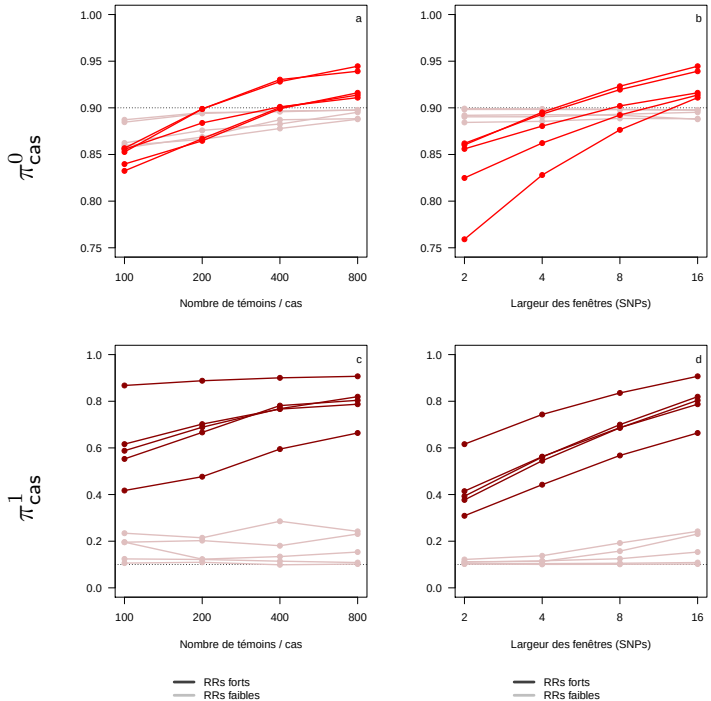


Figure 3.13 Taux de succès périTIMs des cas primitifs (a,b) et mutants (c,d) en fonction de la taille de l'échantillon (a,c) et de la largeur des fenêtres (b,d), par risques relatifs. Chaque point représente la moyenne d'un taux de succès périTIM sur les 100 populations. La ligne pointillée représente le taux de succès aléatoire (a,b : 0,9 ; c,d : 0,1).
a,c : fenêtres de 16 SNPs ; b,d : 400/400 témoins/cas.

3.3.5 Effet sur MapARG

Nous avons vu que le taux de succès global π et ses 4 taux de succès partiels π^0_{tem}, π^1_{tem}, π^0_{cas} et π^0_{cas} forment un pic positif autour du TIM, et que ce pic, mesuré par les taux périTIMs, est significativement affecté par nos 4 facteurs testés. Cependant, contre toute attente, le taux de succès global utilitaire $\pi_{\mathsf{utilitaire}}$, que nous supposons représenter mieux la justesse des données réellement utilisées par MapARG (section 3.3.1), ne montre aucun pic au TIM.

Il a été observé précédemment que des haplotypes mutants au TIM, erronément supposés primitifs, étaient beaucoup moins dommageables à MapARG que des haplotypes primitifs erronément assumés mutants (Vahey, 2008). L'une des raisons avancées est que l'arbre généalogique partiel des mutants d'un échantillon stratifié est beaucoup plus petit que celui des primitifs, puisque les mutants auraient en commun de porter une mutation apparue relativement récemment. Ainsi, on pourrait s'attendre à ce que les haplotypes mutants d'un tel échantillon soient plus homogènes que les primitifs. Un haplotype primitif erronément inféré

mutant contaminerait donc de façon plus dramatique la généalogie partielle des mutants que l'absence dans cette généalogie partielle d'un haplotype mutant considéré primitif.

Dans cette perspective, le taux de succès global utilitaire $\pi_{\text{utilitaire}}$ pourrait ne pas être le plus représentatif de ce qui est réellement important pour le bon fonctionnement de la méthode MapARG. Si le plus important est que les primitifs soient le moins souvent possible erronément estimés comme mutants, il faudrait alors se pencher sur le taux de succès semi-partiel π^0.

La figure 3.14 montre l'effet des RRs combinés (*a*), de la taille de l'échantillon (*b*) et de la largeur des fenêtres (*c*) sur le taux de succès semi-partiel des primitifs π^0. Le taux π^0 se comporte similairement à celui des cas primitifs π^0_{cas} (figure 3.9, *a*,*b*,*c*). Pour des RRs faibles, il est plat sur toute la séquence à hauteur de son taux aléatoire (*a*). Le pic périTIM n'apparait qu'avec des RRs forts, et le taux sur le reste de la séquence se détériore proportionnellement aux RRs. La taille de l'échantillon n'exerce à peu près aucun effet

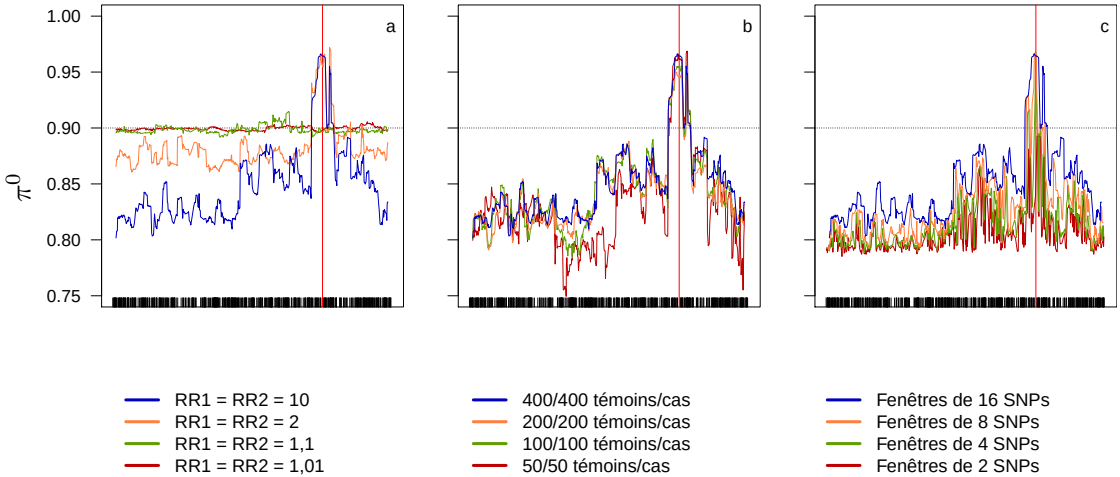


Figure 3.14 Taux de succès des primitifs en fonction des risques relatifs RR1 et RR2 combinés (*a*), de la taille de l'échantillon (*b*) et de la largeur des fenêtres (*c*). L'échelle des ordonnées est la même. La ligne pointillée représente le taux de succès aléatoire (0,9).
a : 400/400 témoins/cas, fenêtres de 16 SNPs ;
b : RR1 = RR2 = 10, fenêtres de 16 SNPs ;
c : RR1 = RR2 = 10, 400/400 témoins/cas.

sur le taux π^0 avec des fenêtres de 16 SNPs (*b*) alors que des fenêtres de plus en plus larges l'améliorent graduellement pour un échantillon de 400/400 témoins/cas (*c*). Les figures analogues pour les 4 taux de succès semi-partiels π^0, π^1, π_{tem} et π_{cas} peuvent être consultées à l'appendice G.

La figure 3.15 montre l'effet des différents facteurs sur le taux de succès périTIM des primitifs obtenu avec 100 populations. On y constate que l'effet positif sur π^0 de la taille des échantillons et de la largeur des fenêtres n'est présent qu'avec des RRs forts, et que de trop courtes fenêtres peuvent même contribuer à détériorer le taux de succès. Les figures analogues pour les 4 taux de succès périTIMs semi-partiels π^0, π^1, π_{tem} et π_{cas} peuvent être consultées à l'appendice H.

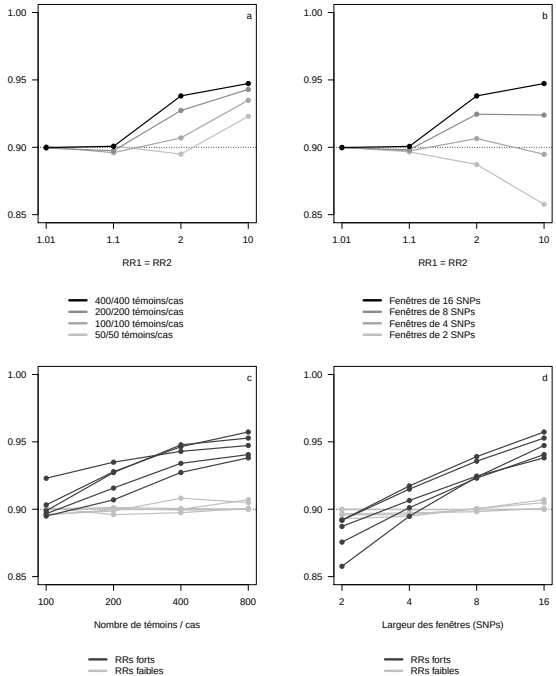


Figure 3.15 Taux de succès périTIMs des primitifs (π^0) en fonction des risques relatifs RR1 et RR2 combinés (*a*,*b*), par taille de l'échantillon (*a*) et par largeur de fenêtre (*b*), et en fonction de la taille de l'échantillon (*c*) et de la largeur des fenêtres (*d*), par risques relatifs. Chaque point représente la moyenne d'un taux de succès périTIM sur les 100 populations. La ligne pointillée représente le taux de succès aléatoire (0,9).
a,*c* : fenêtres de 16 SNPs ; *b*,*d* : 400/400 témoins/cas.

≪ < > ≫

3.4 Discussion

Nous avons décrit au début de ce chapitre une méthode d'estimation des allèles au TIM reposant sur un algorithme EM. Nous avons démontré que la méthode fonctionne relativement bien. Le taux de succès global, ainsi que ses 4 taux partiels et ses 4 taux semi-partiels, ont tous montré un pic évident dans la région du TIM, qui se démarque significativement du taux aléatoire. De plus, les taux périTIMs correspondants ont montré une dépendance par rapport aux facteurs testés. La méthode s'avéra peu efficace en présence de risques relatifs faibles (1,01 et 1,1), mais très efficace avec de forts RRs (2 et 10), particulièrement avec des échantillons suffisamment grands et des fenêtres suffisamment larges. Il peut être décevant de constater que le taux de succès global utilitaire n'ait pas montré de hausse significative autour du TIM, alors qu'on le supposait représenter mieux que les autres taux les distributions ultimement utilisées par MapARG. Il est cependant probable qu'il n'en est rien et que le taux de succès des primitifs soit possiblement plus important.

Cette méthode d'estimation des allèles au TIM repose sur la supposition que nous connaissons le modèle de pénétrance du TIM recherché. Cependant, ce dernier est plus souvent qu'autrement inconnu. Dans cette suite logique et rétroactive, nous testerons au chapitre IV une méthode d'estimation du modèle de pénétrance qui repose sur le même algorithme EM que celui présenté dans ce chapitre.

NOTATIONS DU CHAPITRE III

$d = [h_1,h_2]$	Diplotype d'un individu, où h_1, $h_2 \in \{0,1\}$ représentent ses haplotypes maternel et paternel, respectivement
$d^* = \{d, T\}$	Diplotype d'un individu, incluant ses allèles au TIM
D, Φ	Échantillon aléatoire simple de diplotypes tirés de la population et phénotypes associés
D^*	Échantillon aléatoire simple de diplotypes tirés de la population, incluant le TIM
$\delta \in \{0,1\}$	Allèle d'un haplotype au TIM
f	Fréquence des individus diploïdes dans la population ayant le phénotype $\phi = 1$
$F = \{f_0, f_1, f_2\}$	Modèle de pénétrance
f_i	Probabilité pour un individu d'avoir le phénotype $\phi = 1$ s'il porte i allèles 1 (mutant) au TIM

$m_{h^\delta}^{(k+1)}$	Nombre d'haplotypes de type h porteurs de l'allèle δ au TIM après l'itération k
p	Fréquence des haplotypes dans la population portant l'allèle $\delta = 1$ au TIM
$\phi \in \{0,1\}$	Phénotype d'un individu diploïde
$T = \delta_1\delta_2$	Diplotype d'un individu au TIM, où δ_1, $\delta_2 \in \{0,1\}$ représentent l'allèle au TIM sur ses haplotypes maternel et paternel, respectivement
V_δ	Distribution des haplotypes porteurs de l'allèle δ au TIM
$V_\delta(h)$	Fréquence des haplotypes de type h parmi ceux qui portent l'allèle δ au TIM

CHAPITRE IV

ESTIMATION DU MODÈLE DE PÉNÉTRANCE

4.1 Problématique

Nous avons testé au chapitre III une méthode d'estimation de l'allèle au TIM porté par chacun des haplotypes d'un échantillon. Cette méthode s'avéra efficace en général, et particulièrement effective sous certaines conditions. Cependant, elle suppose que nous connaissons le modèle de pénétrance $F = \{f_0, f_1, f_2\}$, souvent inconnu. Nous présenterons dans ce chapitre (section 4.2) une méthode d'estimation de F également développée par Boucher (2009) et qui se base sur le même algorithme EM décrit au chapitre III. Nous évaluerons la performance de cette méthode à la section 4.3, puis nous en observerons l'utilité à travers l'exécution subséquente de la méthode d'estimation de l'allèle au TIM, à la section 4.4.

4.2 Méthode

La présente méthode d'estimation du modèle de pénétrance peut se diviser en trois étapes distinctes. La première étape (section 4.2.1) consiste à définir un ensemble fini ξ de modèles de pénétrance F^i possibles. Une fois cet ensemble ξ défini, la deuxième étape (section 4.2.2) consiste à estimer une distribution Ψ sur cet ensemble ξ de modèles de pénétrance F^i possibles. Finalement, une fois que nous avons estimé cette distribution Ψ, la troisième étape (section 4.2.3) consiste à estimer F par $\hat{F}$, qui peut être obtenu par différentes méthodes. La section 4.2.4 résume les étapes de la méthode, puis la section 4.2.5 permet de visualiser où elle se situe dans le pipeline de MapARG.

4.2.1 Ensemble fini ξ des modèles de pénétrance possibles

Pour construire un ensemble ξ des modèles de pénétrance F^i (triplets dans $[0,\ 1]^3$) possibles, nous définirons les pas $\{q_{f_0},\ q_{f_1},\ q_{f_2}\}$ par lesquels les valeurs possibles de $\{f_0,\ f_1,\ f_2\}$ seront incrémentées. Puisque nous supposons que la présence de 1 ou 2 allèles mutants au TIM augmente la probabilité qu'un individu développe le phénotype $\phi = 1$, nous appliquerons les contraintes logiques $0 \leq f_0^i \leq f_1^i \leq f_2^i \leq 1$ et $f_0^i < f_2^i$. Ainsi, par exemple, si $q_{f_0} = 0{,}01$, alors $f_0 \in \{0,\ 0{,}01\ , 0{,}02, \ldots, 0{,}99,\ 1\}$. De plus, pour $f_0^i = 0{,}05$ et $q_{f_1} = 0{,}01$, on aura alors $f_1 \in \{0{,}05,\ 0{,}06, \ldots, 0{,}99,\ 1\}$, etc. Une autre contrainte importante est l'intervalle de la valeur possible de la fréquence de l'allèle mutant au TIM dans la population (p^i). Il est possible de limiter cet intervalle par notre connaissance *a priori* du TIM recherché. Cette limitation optionnelle peut permettre de restreindre considérablement la taille de ξ et ainsi réduire significativement le temps de calcul, qui s'avère important. Afin de tester la méthode dans une large perspective, nous conserverons un intervalle possible pour p^i de $[0,\ 0{,}25]$.

Seront ainsi éléments de ξ tous les modèles F^i remplissant les contraintes suivantes :

$$\begin{aligned} f_0^i &= 0 + x_0\, q_{f_0}, \quad x_0 \in \mathbb{N}; \\ f_1^i &= f_0 + x_1\, q_{f_1}, \quad x_1 \in \mathbb{N}; \\ f_2^i &= f_1 + x_2\, q_{f_2}, \quad x_2 \in \mathbb{N}; \\ f_0^i &< f_2^i \leq 1, \end{aligned}$$

et

$$0 \leq p^i \leq 0{,}25.$$

4.2.1.1 Fréquence de l'allèle mutant du TIM dans la population

En supposant que l'on connaisse la fréquence f dans la population du phénotype $\phi = 1$, ce qui est plutôt commun, la valeur de p^i peut être calculée pour un modèle $F^i = \{f_0^i,\, f_1^i,\, f_2^i\}$ donné. Du tableau 3.1, on a que

$$f \quad = \quad f_0^i(1-p^i)^2 \;+\; 2f_1^i p^i(1-p^i) \;+\; f_2^i p^{i2},$$

ce qui nous donne l'équation quadratique

$$(f_0^i - 2f_1^i + f_2^i)\, p^{i2} \;-\; 2(f_0^i - f_1^i)\, p^i \;+\; (f_0^i - f) \quad = \quad 0,$$

qui se résout par :

$$p^i \quad = \quad \begin{cases} \frac{f - f_0^i}{2(f_1^i - f_0^i)}, & \text{si } f_0^i - 2f_1^i + f_2^i \;=\; 0 \\ \frac{f_0^i - f_1^i \pm \sqrt{f_1^{i\,2} - f_0^i f_2^i + f(f_0^i - 2f_1^i + f_2^i)}}{f_0^i - 2f_1^i + f_2^i}, & \text{sinon,} \end{cases}$$

où il existe une et une seule solution pour laquelle $0 < p^i < 1$.

Rappelons que le vrai modèle de pénétrance F du TIM d'un échantillon détermine la fréquence f des phénotypes $\phi = 1$ dans la population, que l'on suppose connue. Ainsi, pour un échantillon de modèle de pénétrance $F = \{0{,}01,\ f_1,\ f_2\}$, comme ceux utilisés au chapitre III et dans ce chapitre, les valeurs possibles pour f_0^i sont très limitées. En effet, puisque f est fixé, connu et utilisé pour calculer p^i, alors f_0^i, incrémenté par q_{f_0}, atteint rapidement une valeur pour laquelle $p^i > 0{,}25$. Pour cette raison, nous n'estimerons pas f_0, que nous supposerons connu (et égal à 0,01 pour tous nos échantillons).

4.2.2 Distribution Ψ : distance entre les distributions V_0 et V_1

Une fois construit l'ensemble ξ des modèles de pénétrance possibles, la méthode procède à estimer une distribution Ψ sur ces modèles F^i. Cette distribution Ψ sera basée sur une distance calculée entre les distributions V_0 et V_1 obtenues par l'algorithme EM présenté au chapitre III. L'hypothèse est que plus le modèle estimé $\hat{F}$ sera proche du vrai modèle F,

mieux il permettra de différencier les haplotypes mutants au TIM des primitifs. Ainsi, par construction, plus un modèle F^i résultera en une grande distance entre V_0^i et V_1^i, plus son poids sera important dans la distribution Ψ.

Donc, pour chaque $F^i \in \xi$, l'algorithme EM est d'abord utilisé pour estimer V_0^i et V_1^i. Ensuite, la distance $\Psi(F^i)$ entre ces deux distributions est calculée à l'aide de $m^*_{h^0}$ et $m^*_{h^1}$, soient les nombres d'haplotypes de type h porteurs respectivement de l'allèle 0 et 1 au TIM après la dernière itération de l'algorithme EM, ou plus simplement V_0 et V_1, avant normalisation (équation 3.2). Cette distance $\Psi(F^i)$ est simplement la somme des carrés des distances entre $m^*_{h^0}$ et $m^*_{h^1}$ pour tous les types d'haplotypes :

$$\Psi(F^i) \quad = \quad \sum_h (m^*_{h^0} - m^*_{h^1})^2. \tag{4.1}$$

4.2.3 Utilisation de la distribution Ψ pour estimer F

Cette distribution Ψ peut ensuite être utilisée de multiples façons pour estimer F. Nous en proposerons trois. La première, la plus simple, consiste à prendre comme estimateur de F ($\hat{F}$) le F^i qui produit la plus grande distance entre V_0 et V_1, c'est-à-dire :

$$\hat{F}_{\mathsf{max}} \quad = \quad F^i \qquad \text{tel que} \qquad \Psi(F^i) \;\geq\; \Psi(F^j), \qquad \forall j \neq i.$$

Une deuxième méthode d'estimation de F à partir de la distribution Ψ consiste à calculer l'espérance de F. Ainsi, $\hat{F}_{\mathbb{E}} = \{\hat{f}_{0_{\mathbb{E}}},\, \hat{f}_{1_{\mathbb{E}}},\, \hat{f}_{2_{\mathbb{E}}},\}$, où

$$\begin{aligned}
\hat{f}_{0_{\mathbb{E}}} \quad &= \quad \mathbb{E}\left[\hat{f}_0\right] \quad = \quad \frac{1}{\sum_{F^i \in \xi} \Psi(F^i)} \sum_{F^i \in \xi} \Psi(F^i)\, f_0^i; \\
\hat{f}_{1_{\mathbb{E}}} \quad &= \quad \mathbb{E}\left[\hat{f}_1\right] \quad = \quad \frac{1}{\sum_{F^i \in \xi} \Psi(F^i)} \sum_{F^i \in \xi} \Psi(F^i)\, f_1^i; \\
\hat{f}_{2_{\mathbb{E}}} \quad &= \quad \mathbb{E}\left[\hat{f}_2\right] \quad = \quad \frac{1}{\sum_{F^i \in \xi} \Psi(F^i)} \sum_{F^i \in \xi} \Psi(F^i)\, f_2^i.
\end{aligned}$$

Enfin, la troisième méthode proposée consiste également à calculer l'espérance de F sur la distribution Ψ, mais en aggravant l'écart de poids entre les modèles faibles et les modèles forts par une puissance de 8. On a donc $\hat{F}_{\mathbb{E}_8} = \{\hat{f}_{0_{\mathbb{E}_8}}, \hat{f}_{1_{\mathbb{E}_8}}, \hat{f}_{2_{\mathbb{E}_8}}, \}$, où

$$\begin{aligned}
\hat{f}_{0_{\mathbb{E}_8}} &= \mathbb{E}\left[\hat{f}_0\right] = \frac{1}{\sum_{F^i \in \xi} \Psi(F^i)^8} \sum_{F^i \in \xi} \Psi(F^i)^8 \, f_0^i; \\
\hat{f}_{1_{\mathbb{E}_8}} &= \mathbb{E}\left[\hat{f}_1\right] = \frac{1}{\sum_{F^i \in \xi} \Psi(F^i)^8} \sum_{F^i \in \xi} \Psi(F^i)^8 \, f_1^i; \\
\hat{f}_{2_{\mathbb{E}_8}} &= \mathbb{E}\left[\hat{f}_2\right] = \frac{1}{\sum_{F^i \in \xi} \Psi(F^i)^8} \sum_{F^i \in \xi} \Psi(F^i)^8 \, f_2^i.
\end{aligned}$$

Les résultats que nous obtiendront lors de l'évaluation de la méthode d'estimation du modèle de pénétrance nous donneront un aperçu des différences d'efficacité de ces trois utilisations de la distribution Ψ pour estimer F.

4.2.4 Algorithme

Afin de bien saisir la structure de la méthode d'estimation du modèle de pénétrance que nous venons de présenter, en voici les grandes étapes, ainsi que l'insertion de l'algorithme EM dans cette méthode :

I. Déterminer l'ensemble ξ des modèles de pénétrance $F^i = \{f_0^i, f_1^i, f_2^i\}$ possibles, selon les pas $\{q_{f_0}, q_{f_1}, q_{f_2}\}$ et tel que $0 \leq f_0^i \leq f_1^i \leq f_2^i \leq 1$ et $0 \leq p^i \leq 0{,}25$;

II. Pour chacun des modèles de pénétrance $F^i \in \xi$:

   i. **Estimer les distributions V_0^i et V_1^i avec l'algorithme EM** (section 3.2.4) ;

   ii. Calculer la distance $\Psi(F^i)$ entre V_0^i et V_1^i (équation 4.1) ;

III. Estimer le modèle de pénétrance F par l'une des trois méthodes proposées :

$$\hat{F} \quad = \quad \begin{cases} \hat{F}_{\mathsf{max}}, & \text{ou bien} \\ \hat{F}_{\mathbb{E}}, & \text{ou bien} \\ \hat{F}_{\mathbb{E}_8} & . \end{cases}$$

4.2.5 Implantation dans MapARG

Dans l'optique de la vraisemblance composite, la présente méthode d'estimation du modèle de pénétrance est utilisée à chaque fenêtre, avant l'utilisation de la méthode du chapitre III pour estimer les allèles au TIM. Ainsi, pour chaque fenêtre w, une distribution Ψ^w des modèles de pénétrances est estimée, ainsi qu'un modèle de pénétrance $\hat{F}^w$, subséquemment utilisé pour estimer les allèles au TIM des haplotypes, dans cette fenêtre w. Voici où s'insère cette méthode d'estimation du modèle de pénétrance dans le pipeline de MapARG :

I. Choisir l'ensemble des positions x_T pour lesquelles $CL(x_T)$ sera évaluée ;
II. Pour chacune des $L - d + 1$ fenêtres couvrant l'ensemble des SNPs de l'échantillon :
    1. **Estimer la distribution Ψ^w et un modèle de pénétrance $\hat{F}^w$** (section 4.2.4) ;
    2. Estimer les distributions V_0 et V_1 par l'algorithme EM avec $\hat{F}^w$ (section 3.2.4) ;
    3. Pour chacun des $d - 1$ intervalles situés dans la fenêtre :

        Pour chacun des K graphes à construire :

1. Pour chaque individu de l'échantillon :
   i. Générer, avec $\hat{V}_0$, $\hat{V}_1$, les équations 3.4.a à 3.4.h, et par l'équation 3.6, la distribution de ses 4 T possibles ;
   ii. Générer $T = \delta_1\delta_2$ selon cette distribution ;
2. Pour chaque étape τ du graphe, tant que le MRCA n'est pas atteint :
   i. Calculer $\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{P_{x_T}(H_{\tau+1}|H_\tau)} = \frac{\phi(H_\tau)}{\phi(H_{\tau+1})}$;
   ii. Mettre à jour $Q_{x_T}(H_\tau)$ et $P_{x_T}(H_{\tau+1})$;
   iii. Générer un évènement selon $P_{x_T}(H_{\tau+1})$;
   iv. Mettre à jour $H_{\tau+1}$;

III. Pour chaque position x_T :

Calculer $\hat{CL}(x_T)$;

IV. $\hat{x}_T$ correspond au maximum de $\hat{CL}(x_T)$.

4.3 Évaluation de la méthode

Tout comme pour l'évaluation de la méthode d'estimation de l'allèle au TIM au chapitre III, la présente méthode d'estimation du modèle de pénétrance sera évaluée fenêtre par fenêtre le long de la séquence. Afin d'évaluer l'efficacité de la méthode à bien estimer le modèle de pénétrance, une distance euclidienne sera calculée pour chacune des trois méthodes d'estimation de F, entre le vrai modèle de pénétrance F du TIM et le modèle estimé $\hat{F}^w$ de la fenêtre w. Puisque nous supposons connu le f_0 du réel modèle F, les 3 distances euclidiennes, normalisées sur $[0,1]$, ne seront obtenues que par les distances avec f_1 et f_2 :

$$\Upsilon^w_{\mathsf{max}} = \frac{1}{\sqrt{2}}\sqrt{(\hat{f}^w_{1_{\mathsf{max}}} - f_1)^2 + (\hat{f}^w_{2_{\mathsf{max}}} - f_2)^2} \quad ;$$

$$\Upsilon^w_{\mathbb{E}} = \frac{1}{\sqrt{2}}\sqrt{(\hat{f}^w_{1_{\mathbb{E}}} - f_1)^2 + (\hat{f}^w_{2_{\mathbb{E}}} - f_2)^2} \quad ;$$

$$\Upsilon^w_{\mathbb{E}_8} = \frac{1}{\sqrt{2}}\sqrt{(\hat{f}^w_{1_{\mathbb{E}_8}} - f_1)^2 + (\hat{f}^w_{2_{\mathbb{E}_8}} - f_2)^2} \quad .$$

Nous utiliserons également, comme statistique, l'espérance des distances $\Psi^w(F^i)$ utilisées par chacune des trois méthodes pour obtenir leurs $\hat{F}^w$ respectifs, c'est-à-dire :

$$\Lambda^w_{\mathsf{max}} = \mathbb{E}\left[\Psi^w(\hat{F}^w_{\mathsf{max}})\right] = \Psi^w(\hat{F}^w_{\mathsf{max}}) \quad ;$$

$$\Lambda^w_{\mathbb{E}} = \mathbb{E}\left[\Psi^w(\hat{F}^w_{\mathbb{E}})\right] = \frac{1}{\gamma(\xi)} \sum_{F^i \in \xi} \Psi^w(F^i) \quad ;$$

$$\Lambda^w_{\mathbb{E}_8} = \mathbb{E}\left[\Psi^w(\hat{F}^w_{\mathbb{E}_8})\right] = \frac{1}{\gamma(\xi)} \sum_{F^i \in \xi} \Psi^w(F^i)^8 \quad ,$$

où $\gamma(\xi)$ est la taille de ξ.

Les mêmes échantillons que ceux de la section 3.3.3 seront utilisés dans ce chapitre, et les mêmes facteurs que ceux de la section 3.3.2 seront testés, soient les risques relatifs, la taille de l'échantillon et la largeur des fenêtres. Les pas utilisés pour générer les modèles $F^i \in \xi$ seront $q_{f_1} = 0{,}01$ et $q_{f_2} = 0{,}01$, ce qui donnera un peu moins de $5\,000$ modèles de pénétrances à tester à chaque fenêtre, dépendamment de la fréquence f des phénotypes $\phi = 1$ associée au vrai modèle de pénétrance F d'un échantillon, qui affectera les fréquences p^i des modèles F^i possibles (voir section 4.2.1.1).

4.3.1 Répartition spatiale des estimations de F par les 3 méthodes

Observons tout d'abord la répartition spatiale des estimations de F par les 3 méthodes, soient $\hat{F}_{\mathbb{E}}$, $\hat{F}_{\mathbb{E}_8}$ et $\hat{F}_{\mathsf{max}}$ pour 4 modèles de pénétrance, avec des échantillons de 400/400 témoins/cas et des fenêtres de 16 SNPs (figure 4.1). Chaque point identifie le modèle estimé $\hat{F}^w$ à la fenêtre w par l'une des 3 méthodes, selon la couleur. La zone grise identifie l'espace

de ξ, et les 2 droites rouges, horizontale et verticale, situent respectivement les f_1 et f_2 du vrai modèle F.

On observe que tous les $\hat{F}_{\mathbb{E}}$ (en bleu), peu importe l'importance des risques relatifs, semblent se concentrer près du centre de masse de l'espace de ξ (non montré), comme si tous les F^i avaient presque le même poids dans la distribution Ψ. Les $\hat{F}_{\mathbb{E}_8}$ (en vert), quant à eux, se répartissent légèrement plus, et certains tombent très près de la valeur réelle de f_1, sauf lorsque les RRs sont très forts, où ils sont néanmoins situés plus près du F réel que les $\hat{F}_{\mathbb{E}}$.

Les $\hat{F}_{\mathsf{max}}$ (en rouge) se comportent de façon tout à fait différente. Avec des RRs très faibles (1,01), ils se concentrent aux limites de l'espace de ξ, particulièrement sur les droites $f_1 = f_2$ et $f_2 = 1$. Quelques $\hat{F}_{\mathsf{max}}$ se trouvent même très près du F réel. Avec des RRs plus forts, les concentrations semblent migrer vers la droite $f_1 = 0$, puis éventuellement, avec des RRs très forts (10), les $\hat{F}_{\mathsf{max}}$ se concentrent tous à proximité du vrai modèle F.

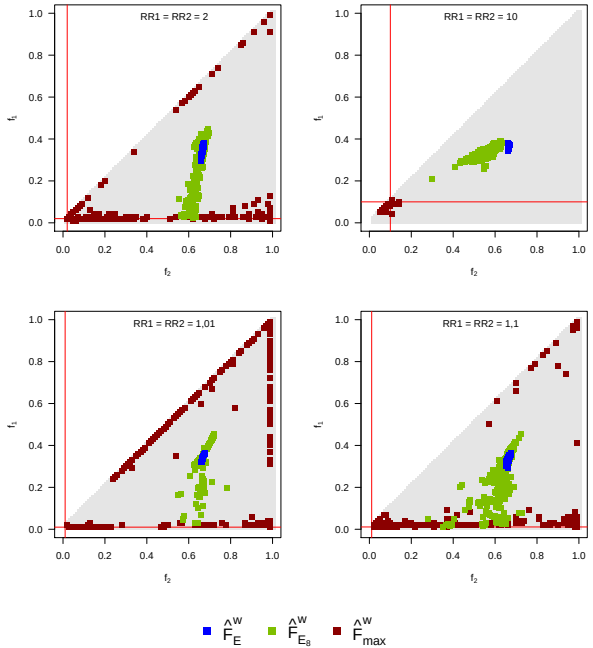


Figure 4.1 Modèle de pénétrance estimé $\hat{F}$ par les 3 méthodes d'estimation, pour 4 modèles de pénétrance (RR1 = RR2 $\in \{1{,}01,\ 1{,}1,\ 2,\ 10\}$), pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs. Chaque point identifie le modèle estimé $\hat{F}^w$ à la fenêtre w par une des 3 méthodes. La zone grise identifie l'espace de ξ. Les droites rouges, horizontale et verticale, situent les f_1 et f_2 du vrai modèle F.

4.3.2 Distance euclidienne Υ du vrai modèle F le long de la séquence

La figure 4.2 permet de visualiser les distances euclidiennes normalisées Υ des 3 méthodes entre le modèle F réel et les estimations $\hat{F}$ le long de la séquence, pour les 4 mêmes échantillons, avec des fenêtres de 16 SNPs. Pour les RRs $=$ 1,01, 1,1 et 2, les distances $\Upsilon_{\mathbb{E}}$ et $\Upsilon_{\mathbb{E}_8}$ sont très stables et autour de 0,5 sur toute la séquence, résultant en une très mauvaise estimation du modèle F. Avec des RRs très forts, les $\hat{F}_{\mathbb{E}}$ et $\hat{F}_{\mathbb{E}_8}$ se rapprochent légèrement de F, toujours indépendamment de la proximité du TIM.

Les Υ_{max} sont beaucoup plus variables. Avec des RRs très faibles (1,01), une forte majorité des $\hat{F}_{\mathsf{max}}$ sont situés plus loin du F que les $\hat{F}_{\mathbb{E}}$ et $\hat{F}_{\mathbb{E}_8}$, et sont même situés tout près de l'extrémité opposée de l'espace de ξ ($\Upsilon_{\mathsf{max}} \to 1$). Néanmoins, une poignée d'entre eux (sur près de 500 fenêtres) se situent très près du F. À mesure que les RRs augmentent, les $\hat{F}_{\mathsf{max}}$ migrent vers F et, avec des RRs très forts (10), se situent tous relativement près de F ($\Upsilon_{\mathsf{max}} \to 0$). On remarque que, avec les RRs $=$ 2, toutes les distances Υ_{max} situées tout

près du TIM sont pratiquement nulles et que, avec les RRs $= 10$, un pic semble de dessiner autour du TIM.

La figure 4.3 permet de voir l'effet de la taille de l'échantillon et de la largeur des fenêtres avec un modèle de pénétrance fort (RR1 $=$ RR2 $= 10$). À première vue, aucune des 3 distances Υ ne semble être significativement améliorée par de plus grands échantillons, quelle que soit la largeur des fenêtres utilisées. Cependant, la largeur des fenêtres montre un effet évident sur $\Upsilon_{\mathbb{E}_8}$ et $\Upsilon_{\max}$ (mais pas $\Upsilon_{\mathbb{E}}$). Les $\hat{F}_{\mathbb{E}_8}$ se rapprochent tranquillement de F avec des fenêtres de plus en plus larges, peu importe la taille de l'échantillon. Les $\hat{F}_{\max}$ montrent un comportement particulier : la majorité sont situés ou bien très loin de F ($\Upsilon_{\max} \approx 0{,}9$), ou bien très près de F ($\Upsilon_{\max} \to 0$). À mesure que les fenêtres s'élargissent, les mauvais $\hat{F}_{\max}$ semblent migrer subitement tout près de F.

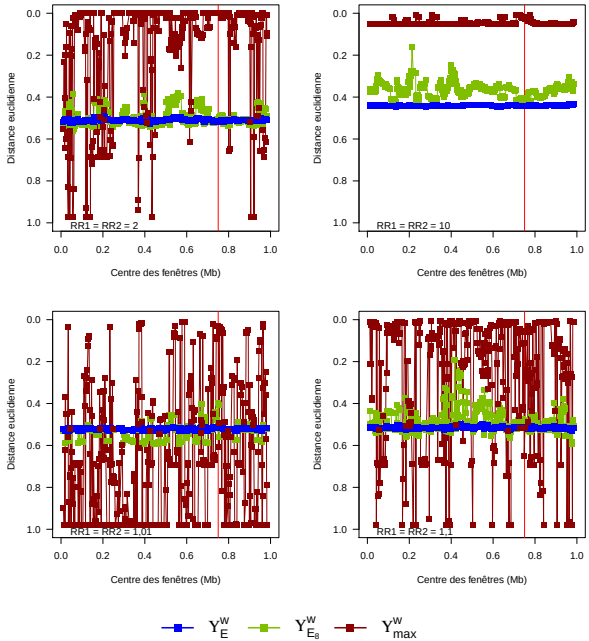


Figure 4.2 Distance euclidienne Υ entre le modèle estimé $\hat{F}$ et le vrai modèle F le long de la séquence pour les 3 méthodes d'estimation, pour 4 modèles de pénétrance (RR1 = RR2 $\in \{1{,}01,\ 1{,}1,\ 2,\ 10\}$), pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs.

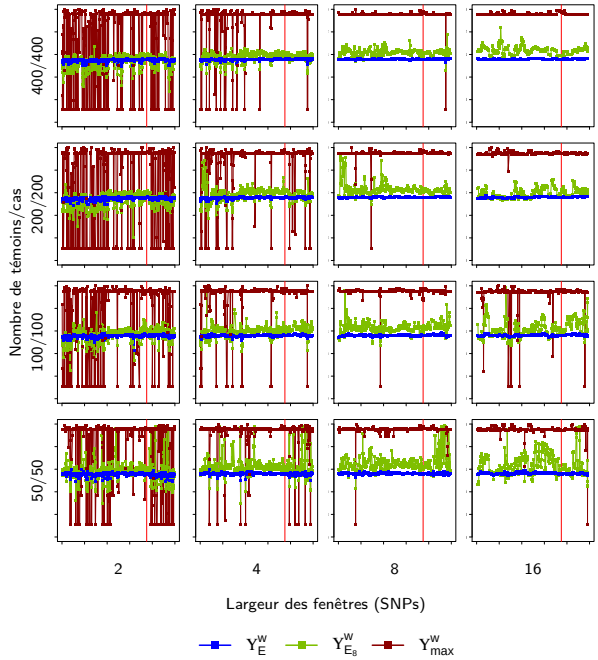


Figure 4.3 Distance euclidienne Υ entre le modèle estimé $\hat{F}$ et le vrai modèle F le long de la séquence pour les 3 méthodes d'estimation, en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = 10 et RR2 = 10. Échelle des ordonnées : [1, 0].

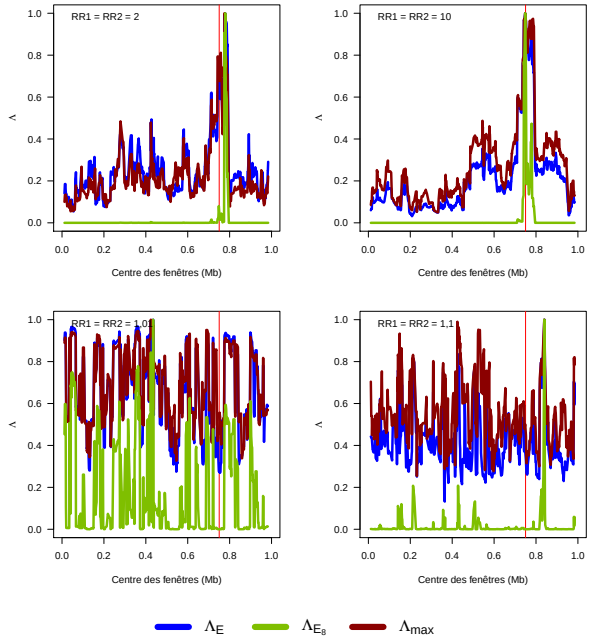


Figure 4.4 Espérance Λ des distances Ψ le long de la séquence pour les 3 méthodes d'estimation, pour 4 modèles de pénétrance (RR1 = RR2 $\in \{1{,}01,\ 1{,}1,\ 2,\ 10\}$), pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs.

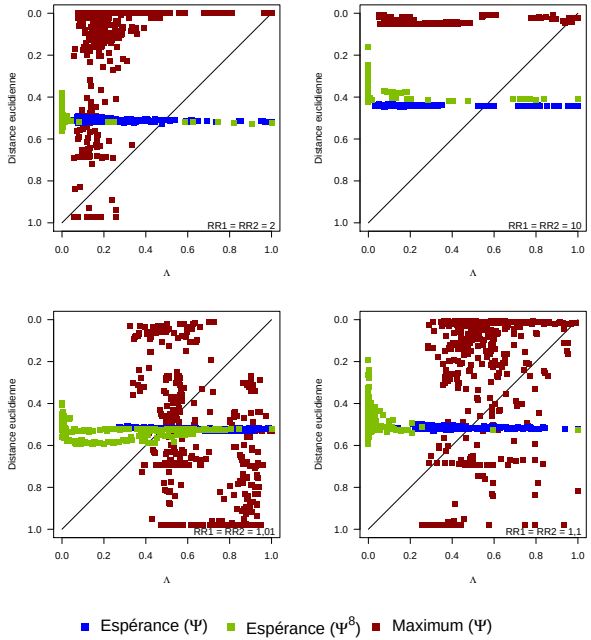


Figure 4.5 Distance euclidienne Υ entre $\hat{F}$ et F en fonction de l'espérance Λ pour les 3 méthodes d'estimation, pour 4 modèles de pénétrance (RR1 = RR2 $\in \{1{,}01,\ 1{,}1,\ 2,\ 10\}$), pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs.

4.3.3 Espérance Λ le long de la séquence

Observons maintenant les espérances $\Lambda_{\mathbb{E}}$, $\Lambda_{\mathbb{E}_8}$ et Λ_{max} qui ont permis d'obtenir les estimés $\hat{F}_{\mathbb{E}}$, $\hat{F}_{\mathbb{E}_8}$ et $\hat{F}_{\mathsf{max}}$, respectivement. La figure 4.4 montre ces espérances le long de la séquence, toujours pour les 4 mêmes échantillons de 400/400 témoins/cas et avec des fenêtres de 16 SNPs. Chacun des 3 Λ est normalisé sur $[0,1]$, 1 correspondant au Λ^w le plus élevé obtenu le long de la séquence, pour chacune des 3 méthodes.

Avec des RRs très faibles (1,01), les 3 Λ montrent peu de variabilité. Avec les RRs $=$ 1,1, $\Lambda_{\mathbb{E}_8}$ montre un pic important situé relativement près du TIM. Cependant, avec des RRs forts (2 et 10), les 3 Λ montrent un fort pic presqu'exactement sur le TIM. Le pic de $\Lambda_{\mathbb{E}_8}$ est d'ailleurs très fortement prononcé. Ces pics sont d'autant plus intéressants qu'ils suggèrent même une potentielle méthode de cartographie génétique fine, basée sur une mesure de distance entre les distributions V_0 et V_1 des haplotypes porteurs des allèles 0 et 1 au TIM, respectivement. Cette méthode, quoique se rapprochant des méthodes conventionnelles

telles que les χ^2 ou les tests exacts de Fisher, utiliserait beaucoup plus d'information, soit celle contenue dans tous les SNPs des haplotypes partiels créés par les fenêtres. Pour cette raison, cette méthode potentielle de cartographie mériterait une éventuelle investigation.

La présente implantation de la méthode d'estimation du modèle de pénétrance implique que l'on estime le modèle F à chaque fenêtre, puis que l'on utilise cette estimation $\hat{F}$ dans cette fenêtre pour estimer les allèles au TIM, etc. Dans une perspective où l'on voudrait d'abord balayer toute la séquence, puis choisir comme estimation globale de F le $\hat{F}^w$ obtenu dans la fenêtre où l'espérance Λ était la plus élevée, il serait avantageux que les espérances Λ élevées soient associées aux estimations $\hat{F}^w$ proches de vrai modèle F, c'est-à-dire aux distances euclidiennes Υ faibles. La figure 4.5 montre la corrélation (ou son absence) entre ces deux variables, pour les 3 méthodes. Idéalement, les points se situeraient près de la diagonale. Particulièrement, pour chacune des 3 méthodes, le point le plus à droite (l'espérance Λ^w la plus élevée le long de la séquence) se situerait le plus haut possible, c'est-à-dire que le $\hat{F}^w$

correspondant serait très près du vrai modèle F. On remarque que, pour les 4 modèles de pénétrance, les $\hat{F}^w_{\mathbb{E}}$ et $\hat{F}^w_{\mathbb{E}_8}$ associés aux $\Lambda^w_{\mathbb{E}}$ et $\Lambda^w_{\mathbb{E}_8}$ respectives les plus élevées sont très loin du vrai modèle F. De son côté, le $\hat{F}^w_{\mathsf{max}}$ associé à l'espérance Λ^w_{max} la plus élevée est aussi très loin de F pour des RRs faibles, mais constitue une excellente estimation de F lorsque les RRs sont forts.

4.3.4 Distribution Ψ^8 abTIM et périTIM des modèles de pénétrance

Il serait intéressant et instructif de pouvoir visualiser une distribution Ψ de tous les modèles de pénétrance $F^i \in \xi$. Pour arriver à cette fin, nous prendrons la distribution Ψ^8, normalisée sur $[0,\, 1]$, et calculerons une distribution Ψ^8 périTIM moyenne avec toutes les fenêtres englobant le TIM, de la même façon qu'à la section 3.3.4. Ainsi, pour tout $F^i \in \xi$, $\Psi(F^i)^8$ périTIM sera égal à la moyenne des $\Psi^w(F^i)^8$ sur les fenêtres w passant sur le TIM. Pour fin de comparaison, nous calculerons également une distribution Ψ^8 *abTIM* éloignée du TIM

sur la séquence, soit autour du $250\,000^{\text{e}}$ nucléotide (situé à 0,25 Mb), de la même manière que pour la distribution Ψ^8 périTIM.

La figure 4.6 montre ces deux distributions (*gauche :* Ψ^8 abTIM ; *droite :* Ψ^8 périTIM) obtenues avec deux échantillons (de 400/400 témoins/cas) de modèles de pénétrance $F = \{0{,}01,\ 0{,}3,\ 0{,}7\}$ (*haut*) et $F = \{0{,}01,\ 0{,}1,\ 0{,}1\}$ (*bas*), à l'aide de fenêtres de 16 SNPs. L'échelle de couleur en bas de la figure permet de comprendre le continuum des valeurs de Ψ^8 représentées. Il est d'abord très surprenant de constater la continuité spatiale des distributions. Étrangement, la distribution Ψ^8 abTIM, construite dans une région dont on suppose l'absence d'effet génétique, semble privilégier les f_1^i et f_2^i les plus près de f_0 (0,01). Autrement dit, les modèles F^i qui résultent en les plus grandes distances entre $\hat{V}_0$ et $\hat{V}_1$ pour les haplotypes partiels dans cette région sont ceux pour lesquels f_1^i et f_2^i sont les plus près de f_0. On remarque aussi que la majorité des modèles $F^i \in \xi$ obtiennent un poids très faible.

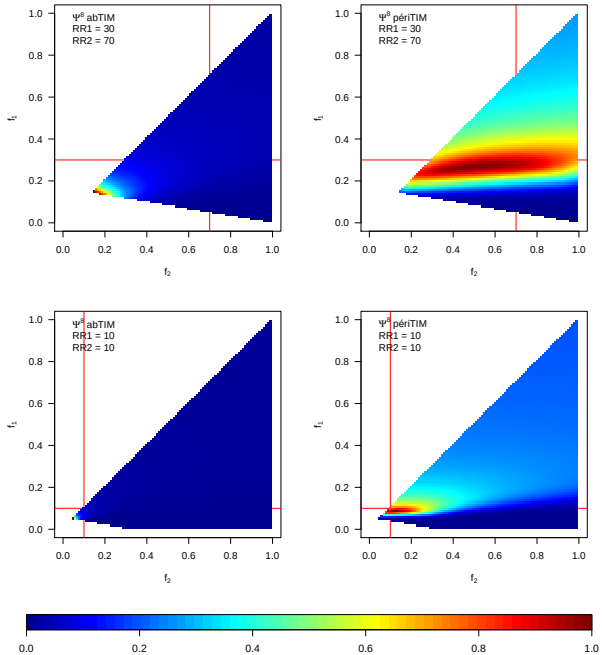


Figure 4.6 Distributions Ψ^8 abTIM (*gauche*) et périTIM (*droite*) normalisées sur $[0,1]$ pour les 2 modèles de pénétrance $\{0{,}01,\ 0{,}3,\ 0{,}7\}$ (*haut*) et $\{0{,}01,\ 0{,}1,\ 0{,}1\}$ (*bas*), pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs. Chaque point situe un $F^i \in \xi$. Les droites rouges, horizontale et verticale, situent les f_1 et f_2 du vrai modèle F.

≪ < > ≫

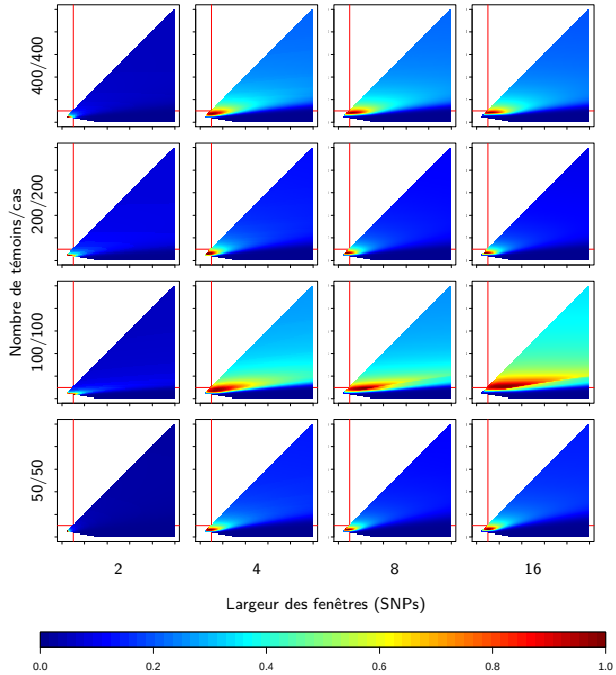


Figure 4.7 Distribution Ψ^8 périTIM normalisée sur $[0,1]$ en fonction de la taille de l'échantillon et des fenêtres, pour RR1 $= 10$ et RR2 $= 10$. Chaque point situe un $F^i \in \xi$. Les droites rouges, horizontale et verticale, situent les f_1 et f_2 du vrai modèle F.

≪ < > ≫

Il est par contre très encourageant de constater que les distributions Ψ^8 périTIM, quant à elles, se concentrent autour du vrai modèle de pénétrance F. On remarque que tous les modèles F^i dont le f_1^i est très faible résultent en une très petite distance $\hat{V}_0 - \hat{V}_1$. Il est aussi intéressant de noter que Ψ^8 périTIM est très concentrée autour du f_1, mais est beaucoup plus éparpillée sur l'axe de f_2. Cette plus grande difficulté à estimer f_2 pourrait être associée à la faible proportion des individus deux fois porteurs de l'allèle mutant au TIM dans un échantillon ($T = 11$), comparativement aux simples porteurs ($T = 01$ et $T = 10$).

La figure 4.7 permet d'observer l'effet de la taille de l'échantillon et de la largeur des fenêtres sur la distribution Ψ^8 périTIM avec le modèle de pénétrance $F = \{0{,}01,\ 0{,}1,\ 0{,}1\}$. On constate que des fenêtres étroites de 2 SNPs donnent de piètres résultats s'apparentant à ceux obtenus dans la région abTIM (non montrés), mais que des fenêtres de 4 SNPs semblent donner une distribution Ψ^8 périTIM tout aussi juste que des fenêtres plus larges. L'effet de la taille de l'échantillon n'est pas très clair, mais semble faire varier la

concentration de Ψ^8.

4.4 Effet sur l'estimation des allèles au TIM

Maintenant que nous avons pu constater le potentiel de la présente méthode à estimer le modèle de pénétrance d'un TIM dans un échantillon, voyons voir si cette estimation $\hat{F}$ du modèle F peut être utilisée efficacement par la méthode d'estimation des allèles au TIM décrite au chapitre III. Pour ce faire, à chaque fenêtre w, le modèle estimé dans cette fenêtre par la méthode du maximum de Ψ^w, soit le modèle $\hat{F}_{\mathsf{max}}$, sera utilisé pour estimer les allèles au TIM, puis les même taux de succès qu'au chapitre III seront calculés.

Les figures 4.8 à 4.15 montrent les 2 taux globaux et les 4 taux partiels obtenus le long de la séquence, par RRs pour des échantillons de 400/400 témoins/cas et des fenêtres de 16 SNPs (4.8 à 4.11), et par taille d'échantillons et largeur de fenêtres pour des RRs de 10 (4.12 à 4.15). Afin de pouvoir bien apprécier l'efficacité de la méthode d'estimation du

modèle de pénétrance, nous comparerons les taux de succès obtenus par les modèles estimés $\hat{F^i}^w$ avec les taux obtenus avec les vrais modèles de pénétrance F. Ainsi, les figures 4.8, 4.9, 4.12 et 4.13 présentent les résultats obtenus par l'estimation fenêtre par fenêtre du modèle, alors que les figures 4.10, 4.11, 4.14 et 4.15 présentent les résultats obtenus en toute connaissance du vrai modèle de pénétrance. La visualisation numérique, deux pages à la fois, est optimale pour fin de comparaison.

Tout d'abord, on constate que les taux de succès globaux sont presque aussi élevés lorsque l'on doit estimer le modèle que lorsqu'on le connait, particulièrement avec des RRs élevés (figures 4.8 et 4.10), où alors le taux global semble même amélioré. Cette légère amélioration du taux global peut être mieux appréciée en comparant les figures 4.12 et 4.14. Indépendamment de la taille de l'échantillon ou de la largeur des fenêtres, on voit que le taux global est légèrement amélioré par l'estimation du modèle de pénétrance, alors que le taux global utilitaire est légèrement diminué. Fait intéressant, le pic au TIM est toujours

bien présent, et la région périTIM semble d'ailleurs être moins être changée par l'estimation du modèle que le reste de la séquence, particulièrement avec des RRs forts.

Cette très forte stabilité de la région périTIM est plus particulièrement visible lorsque l'on observe les taux de succès partiels. En alternant successivement entre les pages 195 et 197, on constate que les pics au TIM des modèles forts ne bougent pratiquement pas, alors que les taux sur le reste de la séquence sont complètement différents avec l'estimation du modèle. De toute évidence, l'estimation du modèle de pénétrance apporte beaucoup plus de variabilité dans les taux de succès. Cette grande variabilité tire souvent les taux vers les extrêmes, parfois les détériorant, mais parfois aussi les amenant à 1, comme c'est le cas pour le taux π^0_{tem} des témoins primitifs. En général, l'estimation du modèle de pénétrance (plutôt que l'utilisation du vrai modèle) semble améliorer le taux des témoins primitifs (π^0_{tem}) et des cas primitifs (π^0_{cas}, particulièrement autour du TIM) et détériorer celui des cas mutants (π^1_{cas}) et plus légèrement celui des témoins mutants (π^1_{tem}).

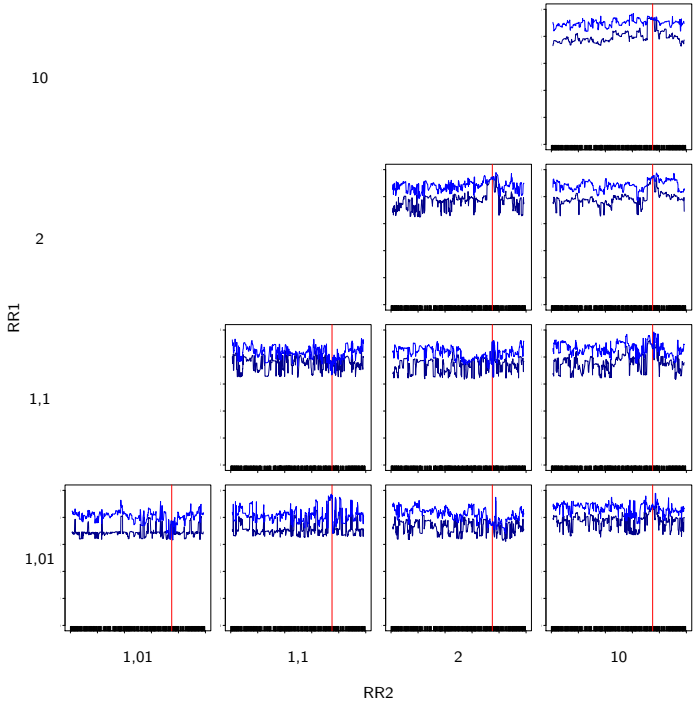


Figure 4.8 [Modèle estimé] Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < > ≫

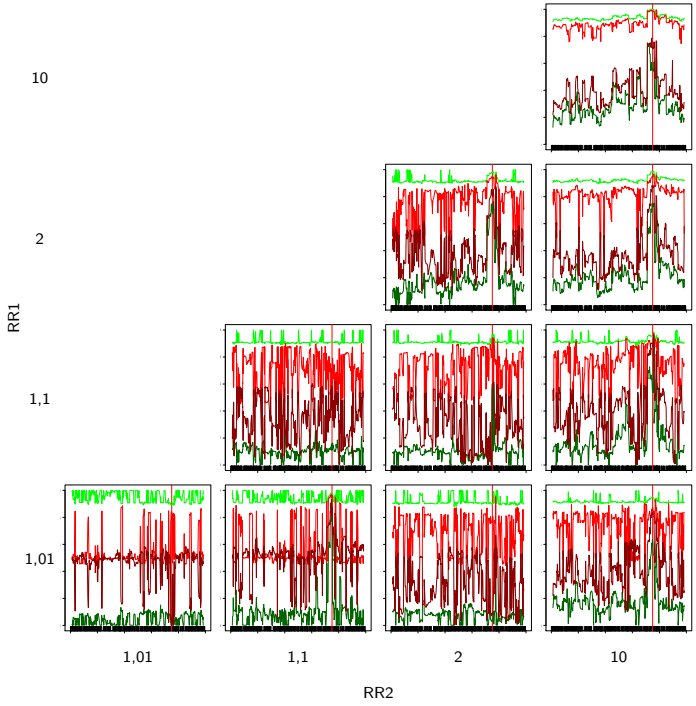


Figure 4.9 [Modèle estimé] Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

≪ < > ≫

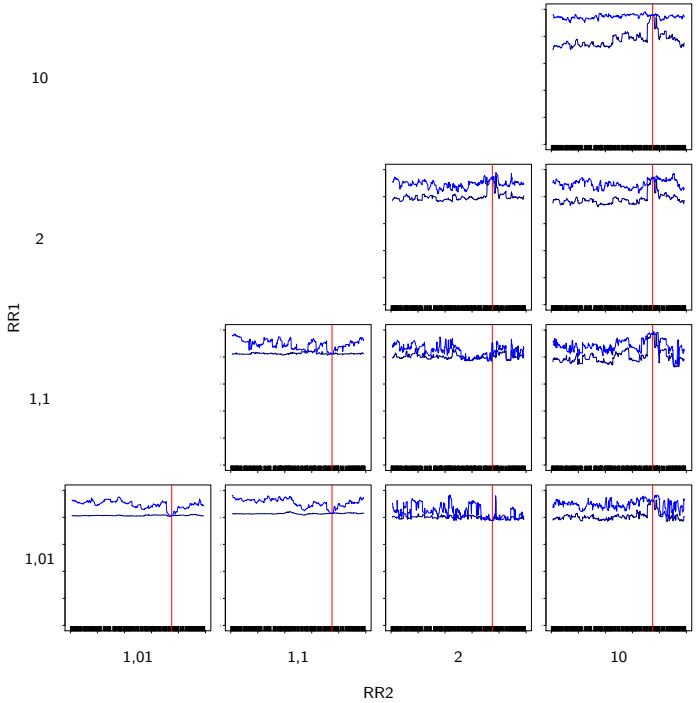


Figure 4.10 [Modèle connu] Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < > ≫

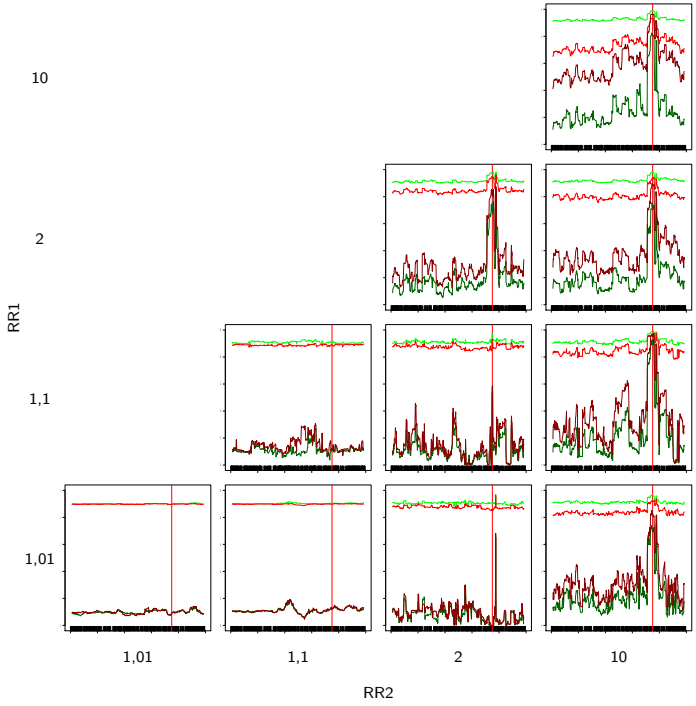


Figure 4.11 [Modèle connu] Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

≪ < > ≫

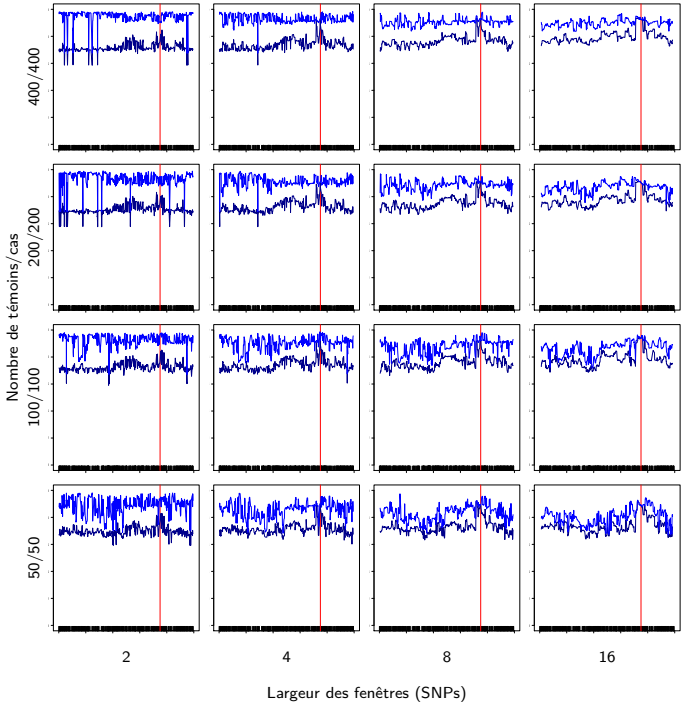


Figure 4.12 [Modèle estimé] Taux de succès globaux en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = 10 et RR2 = 10. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

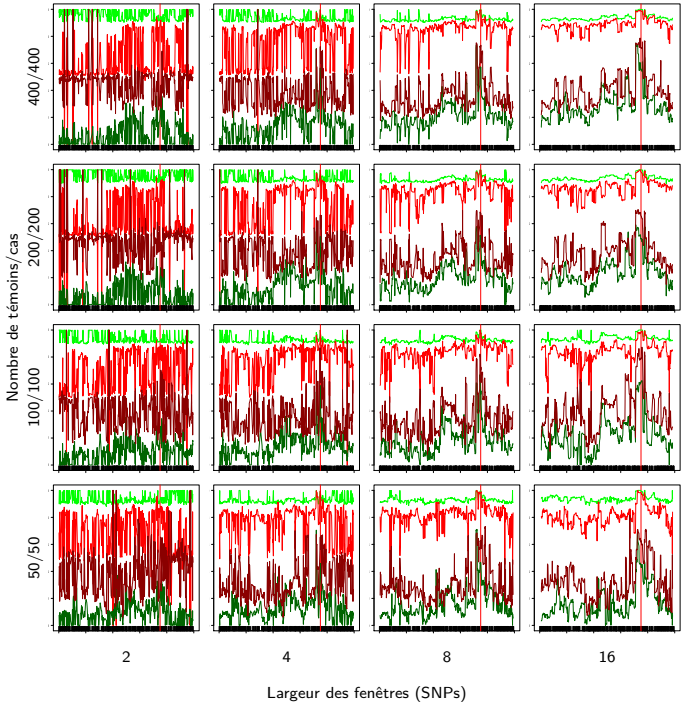


Figure 4.13 [Modèle estimé] Taux de succès partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = 10 et RR2 = 10. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

≪ < > ≫

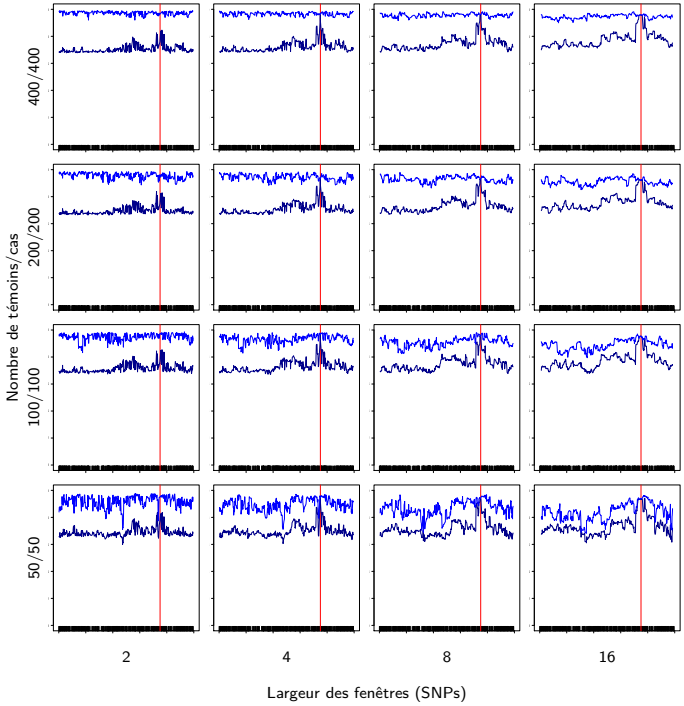


Figure 4.14 [Modèle connu] Taux de succès globaux en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = 10 et RR2 = 10. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

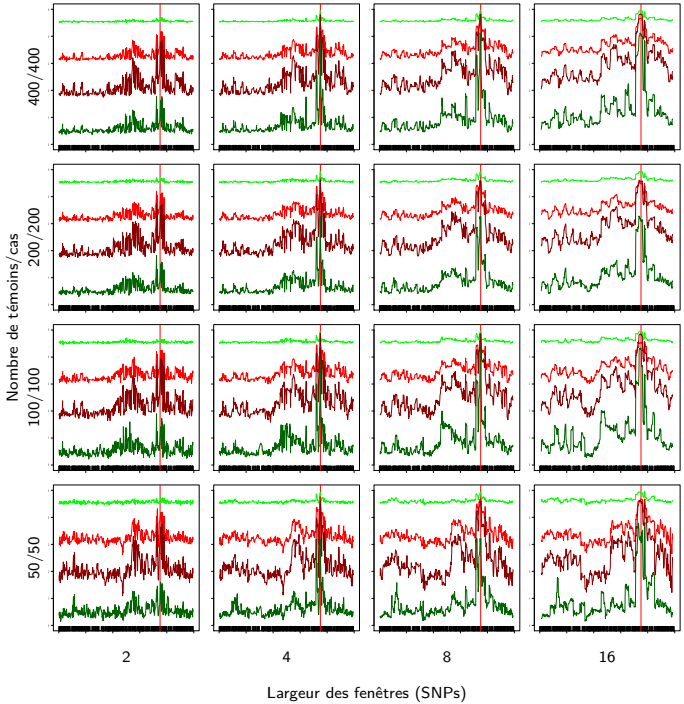


Figure 4.15 [Modèle connu] Taux de succès partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = 10 et RR2 = 10. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

4.5 Discussion

Nous avons présenté dans ce chapitre une méthode d'estimation du modèle de pénétrance qui repose sur le même algorithme EM que la méthode d'estimation des allèles au TIM présentée au chapitre III. La méthode produit une distribution des modèles de pénétrance possibles, et 3 différentes façons d'utiliser cette distribution ont ensuite été proposées pour estimer le modèle. Le potentiel de la méthode a été documenté, et elle mérite une recherche plus approfondie. La possibilité d'une nouvelle méthode de cartographie génétique a d'ailleurs été mentionnée. De plus, l'estimation fenêtre par fenêtre du modèle de pénétrance par cette méthode a montré qu'elle pouvait même améliorer l'estimation subséquente des allèles au TIM, plutôt que l'utilisation du vrai modèle, même s'il est connu.

NOTATIONS DU CHAPITRE IV

f	Fréquence des individus dans la population ayant le phénotype $\phi = 1$
$F = \{f_0,\, f_1,\, f_2\}$	Modèle de pénétrance réel d'un échantillon
F^i	Modèle de pénétrance possible $\in \xi$
$\hat{F}$	Modèle de pénétrance estimé
$\hat{F}^w$	Modèle de pénétrance estimé dans la fenêtre w
$\hat{F}_{\mathbb{E}},\ \hat{F}_{\mathbb{E}_8},\ \hat{F}_{\mathsf{max}}$	Modèles de pénétrance estimés par Ψ, Ψ^8 et le maximum de Ψ
$\gamma(\xi)$	Taille de l'ensemble ξ
$\Lambda_{\mathbb{E}},\ \Lambda_{\mathbb{E}_8},\ \Lambda_{\mathsf{max}}$	Espérances des distances $\Psi(F^i)$, $\Psi(F^i)^8$ et $\Psi(\hat{F}_{\mathsf{max}})$
$m^*_{h^0}$	Nombre estimé d'haplotypes de type h porteurs de l'allèle 0 au TIM
$m^*_{h^1}$	Nombre estimé d'haplotypes de type h porteurs de l'allèle 1 au TIM

p	Fréquence des haplotypes dans la population portant l'allèle 1 au TIM
p^i	Valeur de p correspondant au modèle $F^i \in \xi$
Ψ	Distribution sur ξ basée sur une distance entre V_0 et V_1
$\{q_{f_0}, q_{f_1}, q_{f_2}\}$	Pas par lesquels $\{f_0^i, f_1^i, f_2^i\}$ sont incrémentés pour construire ξ
$\Upsilon_{\mathbb{E}}, \Upsilon_{\mathbb{E}_8}, \Upsilon_{\max}$	Distances euclidiennes, normalisées sur $[0,1]$, entre F et $\hat{F}_{\mathbb{E}}$, $\hat{F}_{\mathbb{E}_8}$ et $\hat{F}_{\max}$
V_0	Distribution des haplotypes porteurs de l'allèle 0 au TIM
V_1	Distribution des haplotypes porteurs de l'allèle 1 au TIM
ξ	Ensemble des modèles de pénétrance F^i possibles

CONCLUSION

L'objectif du présent ouvrage était de tester l'efficacité et le potentiel de deux méthodes d'estimation, l'une pour estimer l'allèle d'une mutation cherchée sur tous les haplotypes d'un échantillon, et l'autre pour estimer le modèle de pénétrance de cette mutation, toutes deux reposant sur le même algorithme EM. Ces deux méthodes d'estimation s'insèrent dans une méthode de cartographie génétique fine que nous avons décrite en détail au chapitre II. La sensibilité de ces méthodes à quatre facteurs fut mise à l'épreuve, soient les risques relatifs RR1 et RR2, la taille des échantillons disponibles ainsi que la largeur des fenêtres utilisées.

En supposant le modèle de pénétrance connu, la méthode d'estimation des allèles s'avère très efficace à bien estimer les témoins et les cas *primitifs* au TIM, peu importe la taille de l'échantillon, la largeur des fenêtres et même pour des RRs très faibles. La juste détection des haplotypes *mutants* (témoins et cas) s'avère toutefois considérablement moins bonne,

et significativement plus sensible aux facteurs testés. En particulier, leurs taux de succès périTIM, très faibles avec des RRs faibles, s'améliorent rapidement avec des RRs forts, particulièrement avec de larges fenêtres. La taille des échantillons ne semble toutefois pas exercer une grande influence. Une bonne estimation des allèles des haplotypes mutants est cependant peut-être moins importante pour la performance subséquente de MapARG que celle des haplotypes primitifs. En effet, la contamination des vrais mutants d'un échantillon, plus homogènes, par des faux primitifs, plus hétérogènes, est possiblement plus dommageable à la méthode de cartographie.

Le modèle de pénétrance étant plus souvent qu'autrement inconnu, nous avons également testé une méthode pour l'estimer, qui produit une distribution sur un ensemble fini discret de modèles de pénétrance possibles. Cette distribution est basée sur une distance résultante entre les haplotypes estimés primitifs et ceux estimés mutants. L'utilisation subséquente de cette distribution pour estimer le vrai modèle peut prendre diverses formes, et nous en avons

décrites trois. La plus performante, aussi la plus variable, est celle consistant à prendre le modèle le plus probable. Elle s'avère cependant peu efficace si les RRs sont très faibles, tout comme les deux autres, qui reposent sur l'espérance de la distribution. Ici encore, la taille des échantillons est peu influente, alors que de larges fenêtres résultent en de bien meilleurs résultats que des petites.

L'estimation subséquente des allèles au TIM en utilisant le modèle estimé (le plus probable) par la méthode fut également comparée à celle utilisant le vrai modèle connu. Il fut d'abord encourageant de constater que les pics des taux de succès autour du TIM étaient toujours présents, et d'intensité quasiment identique. Il fut également observé que les taux de succès des mutants étaient légèrement moins bons qu'avec l'utilisation du vrai modèle, mais que ceux des primitifs étaient plutôt améliorés, particulièrement dans la région périTIM. Si la juste estimation des allèles des primitifs est effectivement plus importante à la méthode MapARG que celle des mutants, l'utilisation du modèle estimé pourrait potentiellement

aider la méthode de cartographie, même si elle se fait au détriment de la bonne estimation des mutants.

La distribution périTIM des modèles de pénétrance, nettement concentrée autour du vrai modèle, au moins pour des RRs forts, comparativement à la distribution abTIM, pourrait s'avérer en soi une méthode de cartographie génétique. En effet, les trois méthodes d'estimation du modèle montrèrent un très fort pic de leur espérance exactement sur la position de la mutation recherchée. Une investigation plus élaborée de ce côté pourrait être prometteuse.

ADN *Acide désoxyribonucléique*. Longue molécule supportant l'information génétique et qui est formée d'une séquence linéaire des nucléotides A, C, G et T. L'ADN humain, long d'environ 3 milliards de nucléotides et divisé en 23 paires de chromosomes, contient environ 30 000 gènes.

ARG *Ancestral Recombination Graph*. Graphe de recombinaison ancestral.

Allèle Une des formes alternatives d'un gène ou d'un nucléotide occupant un locus précis sur un chromosome. Un SNP prend généralement l'un de deux allèles possibles parmi A, C, G et T.

Chromosome Structure macromoléculaire contenant une section précise de la séquence d'ADN. Chez l'humain, l'ADN est divisé en 23 chromosomes. Un individu possède deux copies de chaque chromosome, provenant de chacun de ses deux parents.

Diplotype Ensemble de deux haplotypes homologues d'un individu. Le diplotype, contrairement au génotype, permet de distinguer les deux haplotypes.

Gène Séquence d'ADN située à un locus précis sur un chromosome, souvent longue de plusieurs milliers de nucléotides et codant généralement pour une protéine (définition simplifiée). L'humain possède deux copies de chacun des ses quelques 30 000 gènes.

Génome L'ensemble de toute l'information génétique d'un organisme. Le génome humain est constitué de son ADN divisé en 23 paires de chromosomes, plus son ADN mitochondrial qui n'est transmis que par la mère.

Génotype Composition allélique d'une sélection de marqueurs, sans différentiation du chromosome sur lequel se situe chacun des deux allèles des marqueurs. Le génotype, contrairement au diplotype, ne permet pas de distinguer les deux haplotypes.

Haplotype Composition allélique d'un seul des deux chromosomes d'une paire, pour une sélection de marqueurs. Deux haplotypes homologues d'un individu constituent un diplotype.

LD *Linkage Disequilibrium*. Déséquilibre de liaison.

MRCA *Most Recent Common Ancestor*. Ancêtre commun le plus récent.

Nucléotide Composant moléculaire de base de l'ADN, prenant l'une de quatre formes: A, C, G et T. L'ADN humain est constitué d'environ 3 milliards de nucléotides, et un gène peut en comprendre plusieurs centaines.

Phénotype État d'un caractère observable chez un organisme vivant (anatomique, morphologique, moléculaire, physiologique, ou éthologique). Peu de phénotypes sont uniquement déterminés par l'allèle du/des gène(s) qui lui sont associés. L'environnement exerce généralement une influence, dont l'importance varie selon le caractère observé.

Ploïdie Nombre d'exemplaires de chacun des chromosomes dans une cellule. L'humain étant diploïde, chacune de ses cellules nucléées contient deux chromosomes homologues pour chacune des 23 paires de chromosomes, l'un hérité de sa mère et l'autre de son père.

Polymorphisme Coexistence dans une population de plusieurs allèles pour un gène ou un nucléotide.

SNP *Single Nucleotide Polymorphism*. Nucléotide, situé à un locus précis sur un chromosome, dont une certaine proportion de la population possède un allèle différent du consensus. Chez l'humain, un SNP avec une fréquence allélique $\geq 1\,\%$ est présent à tous les 100 à 300 nucléotides en moyenne, où 2 SNP sur 3 substituent C avec T.

TIM *Trait Influencing Mutation*. Mutation influençant un caractère (phénotypique).

INDEX

APPENDICE A

TAUX DE SUCCÈS GLOBAUX ET PARTIELS, PAR TAILLE DE L'ÉCHANTILLON

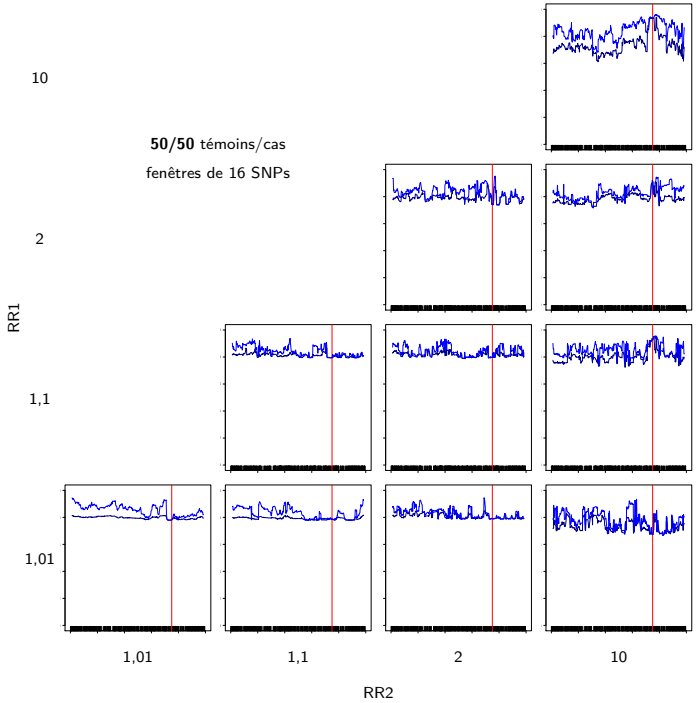


Figure A.I Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de **50/50** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

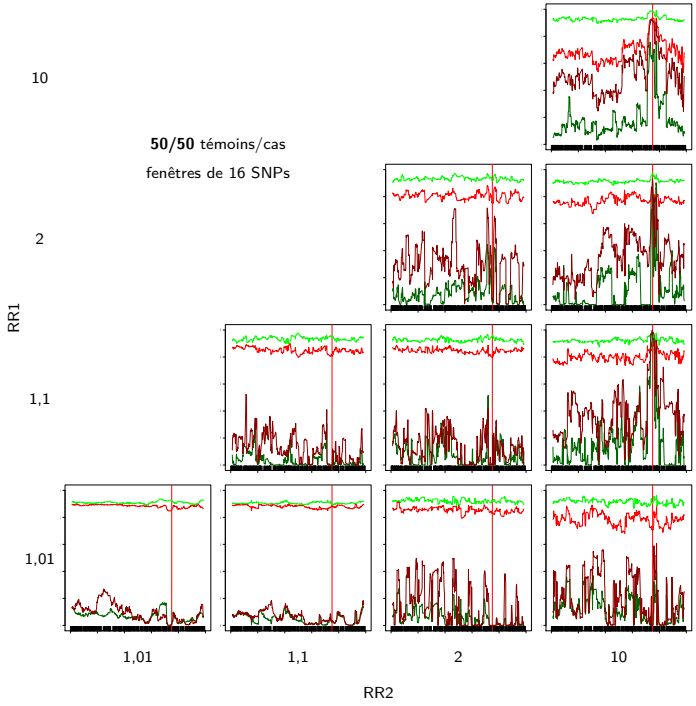


Figure A.II Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **50/50** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

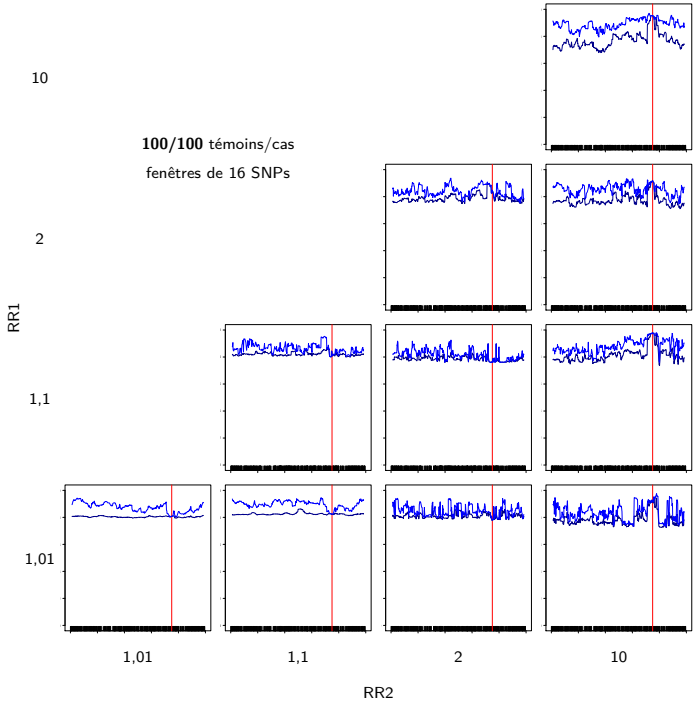


Figure A.III Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de **100/100** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

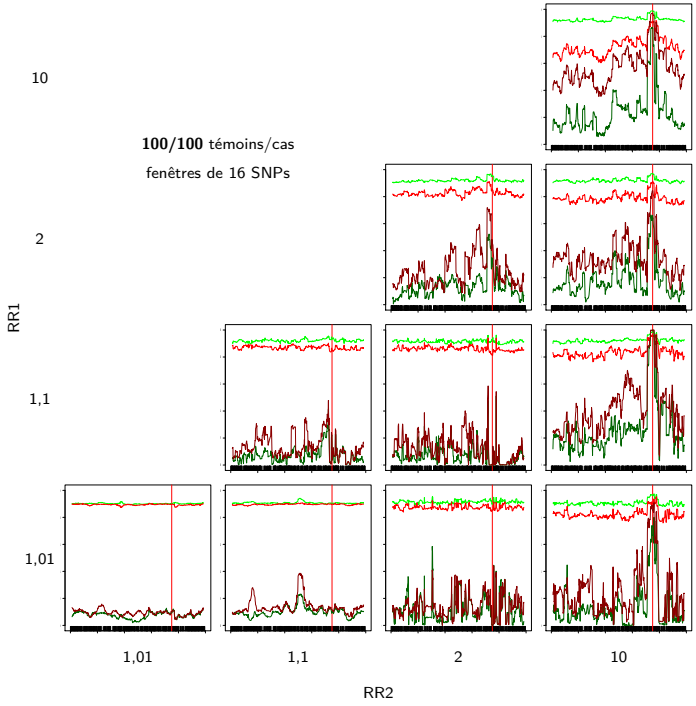


Figure A.IV Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **100/100** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

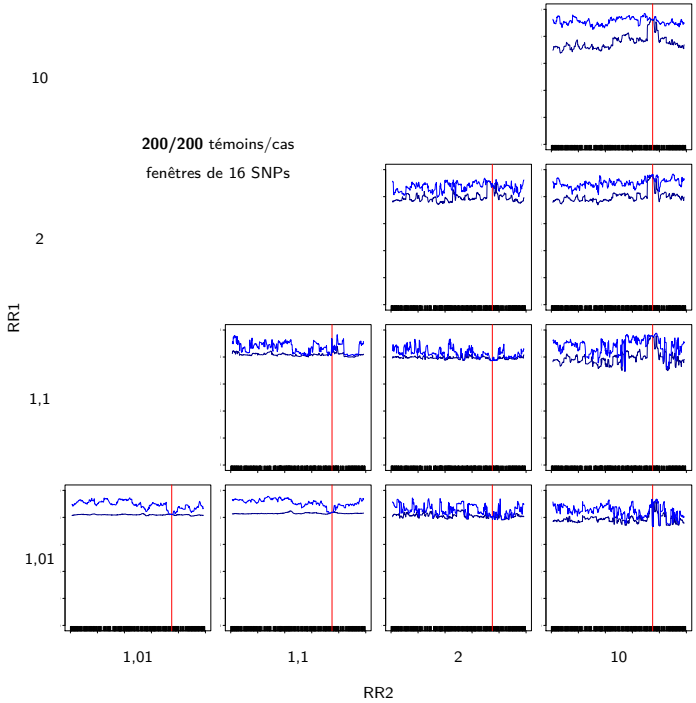


Figure A.V Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de **200/200** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0,\ 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < ≫

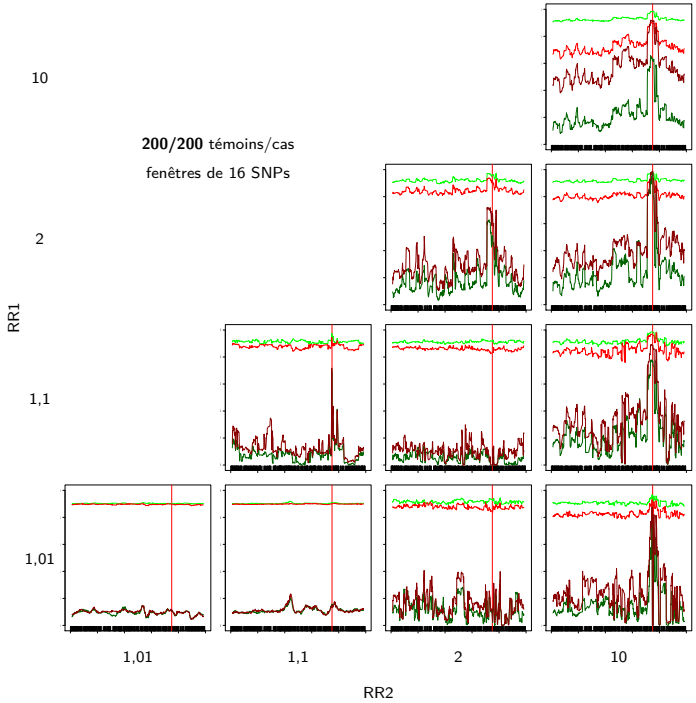


Figure A.VI Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **200/200** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

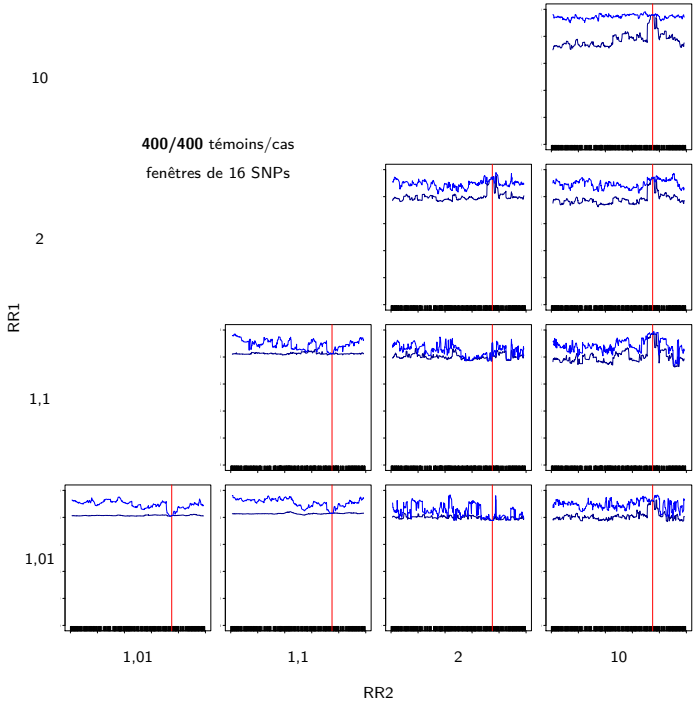


Figure A.VII Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de **400/400** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < ≫

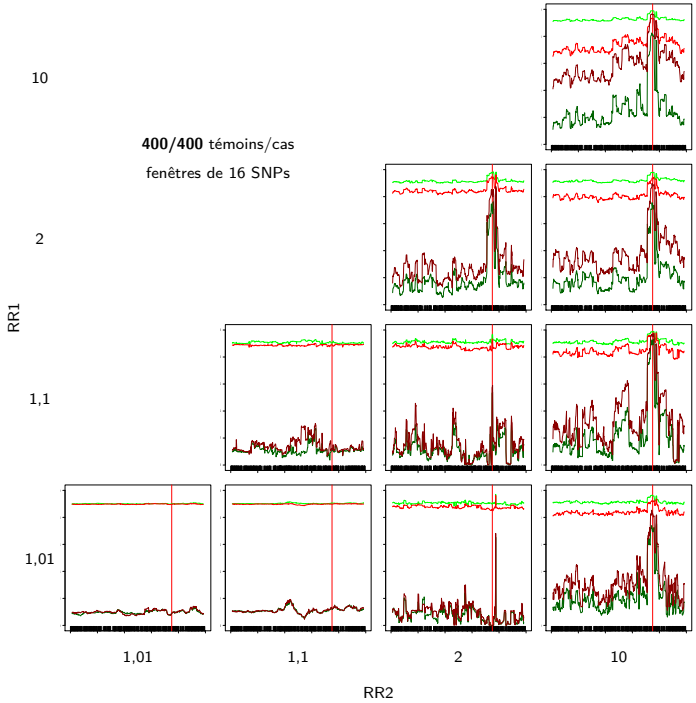


Figure A.VIII Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **400/400** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

APPENDICE B

TAUX DE SUCCÈS GLOBAUX ET PARTIELS, PAR LARGEUR DES FENÊTRES

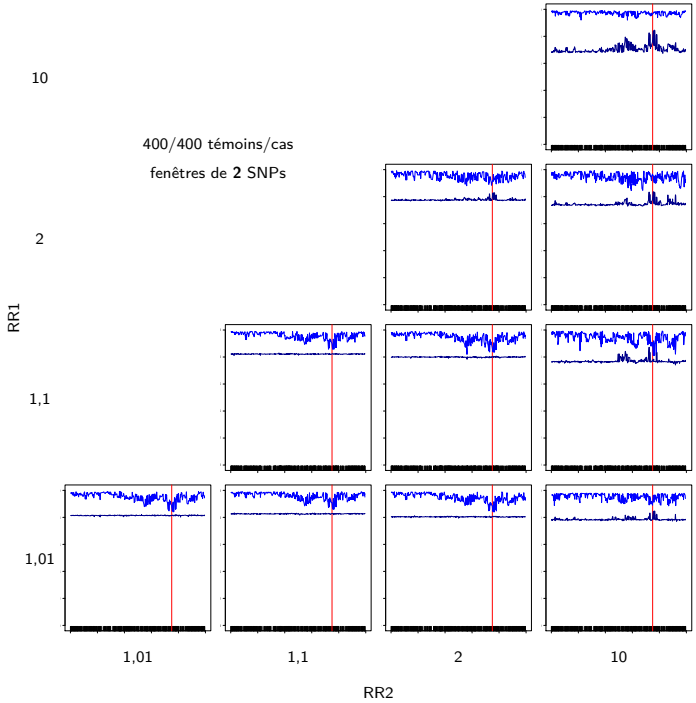


Figure B.I Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **2** SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < ≫

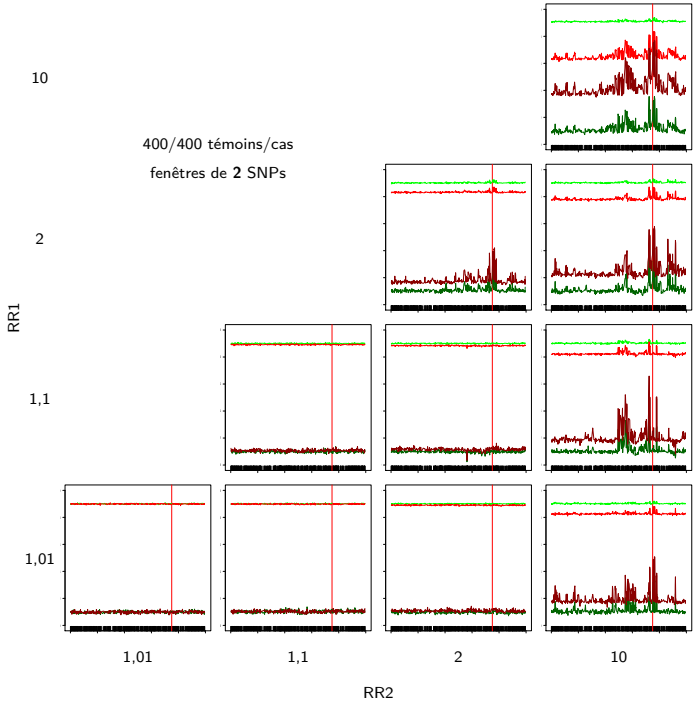


Figure B.II Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **2** SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

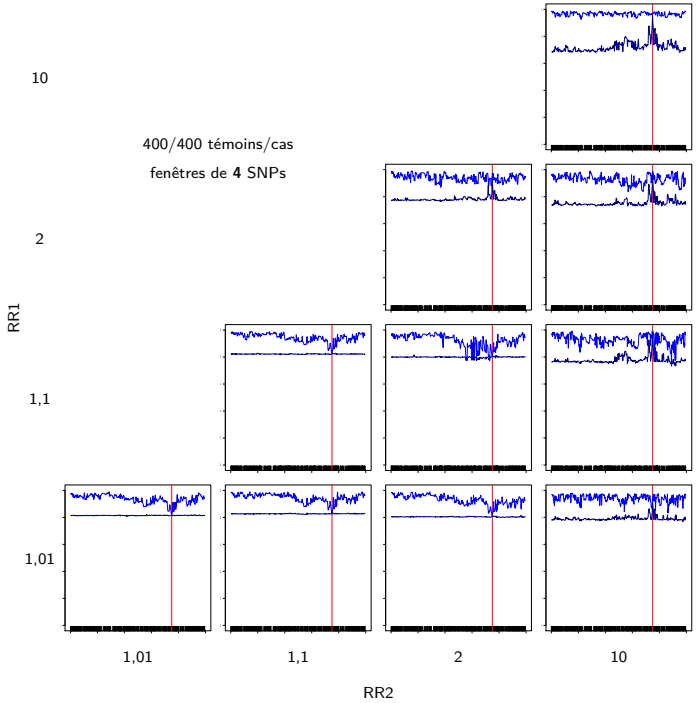


Figure B.III Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **4** SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < ≫

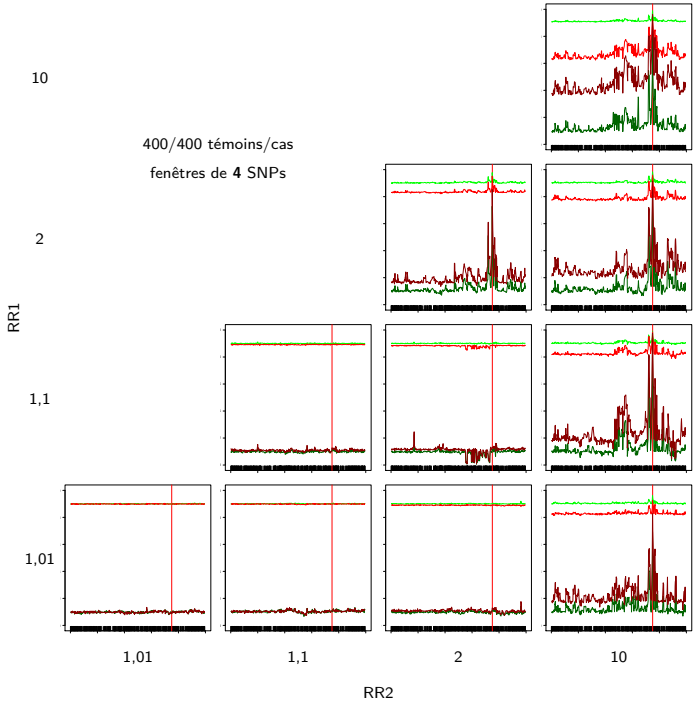


Figure B.IV Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **4** SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

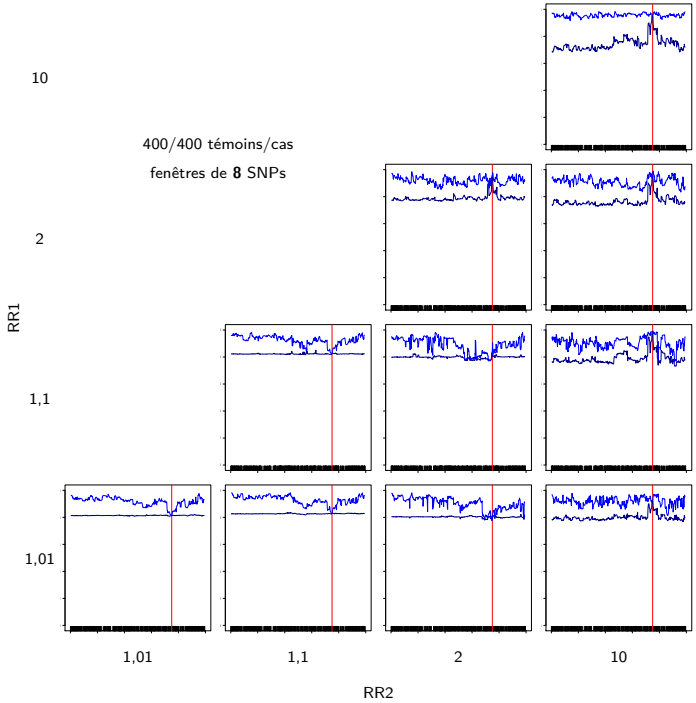


Figure B.V Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **8** SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < ≫

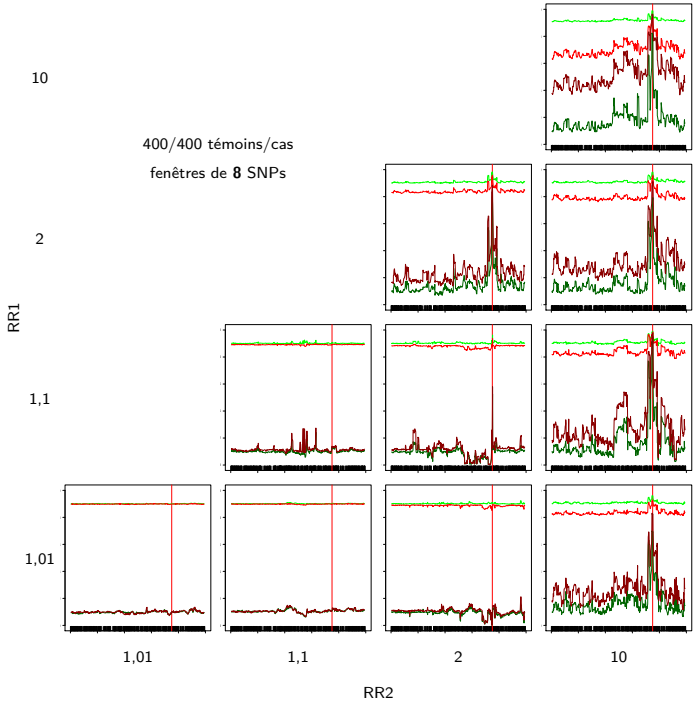


Figure B.VI Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **8** SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

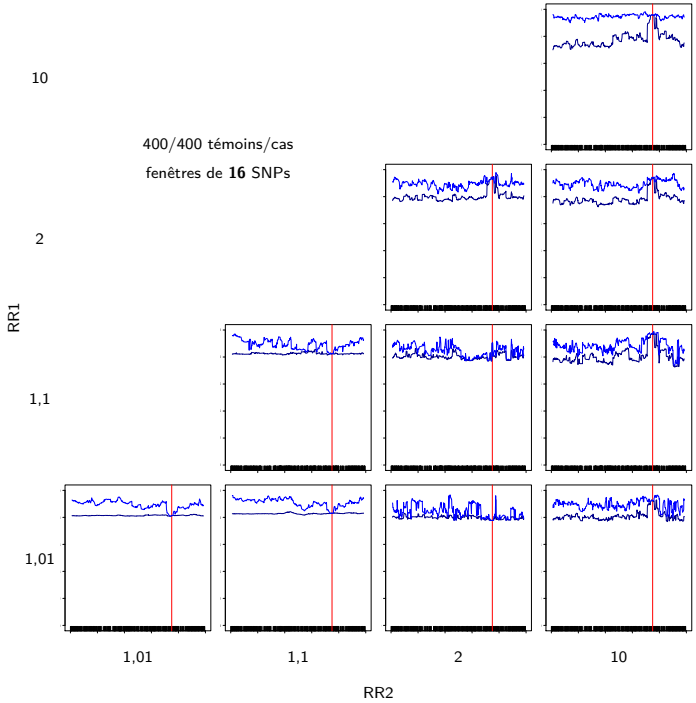


Figure B.VII Taux de succès globaux en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **16** SNPs. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < ≫

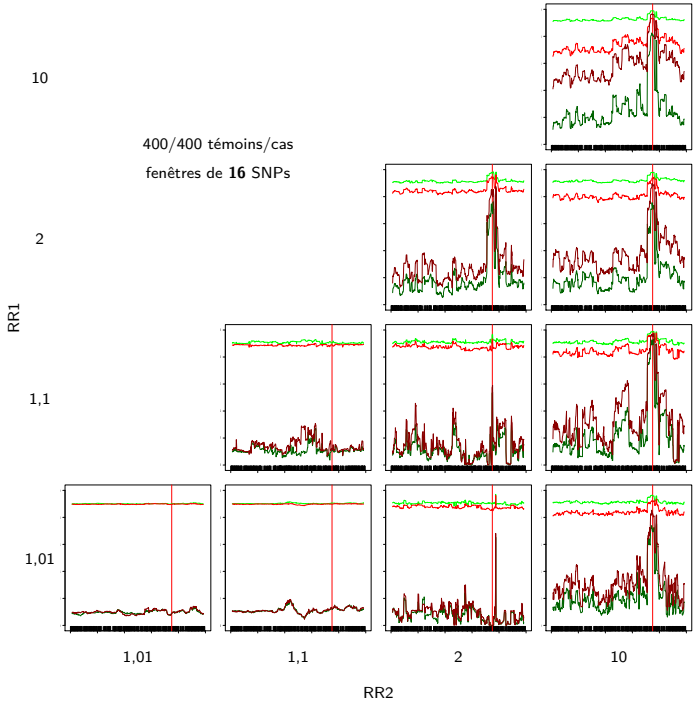


Figure B.VIII Taux de succès partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **16** SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

APPENDICE C

TAUX DE SUCCÈS GLOBAUX ET PARTIELS, PAR RISQUES RELATIFS

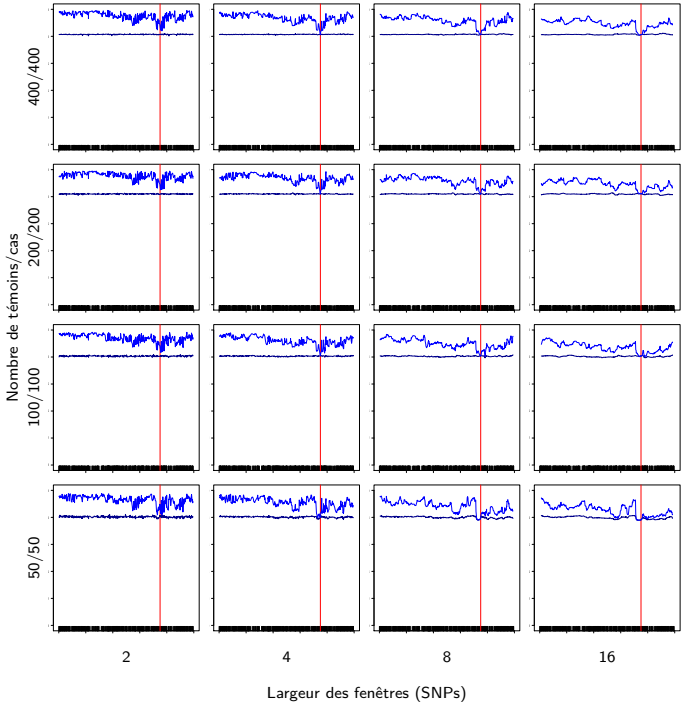


Figure C.I Taux de succès globaux en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **1,01**. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

⋘ < ⋙

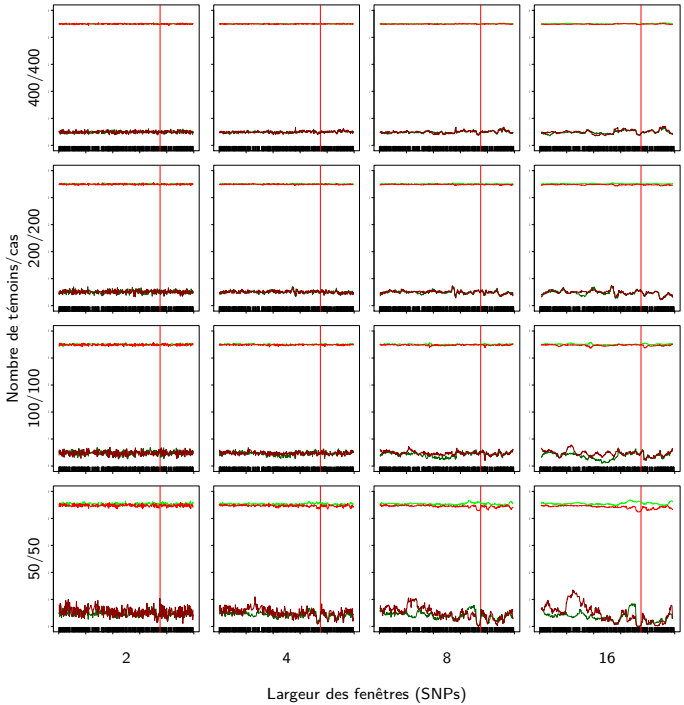


Figure C.II Taux de succès partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **1,01**. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

≪ < ≫

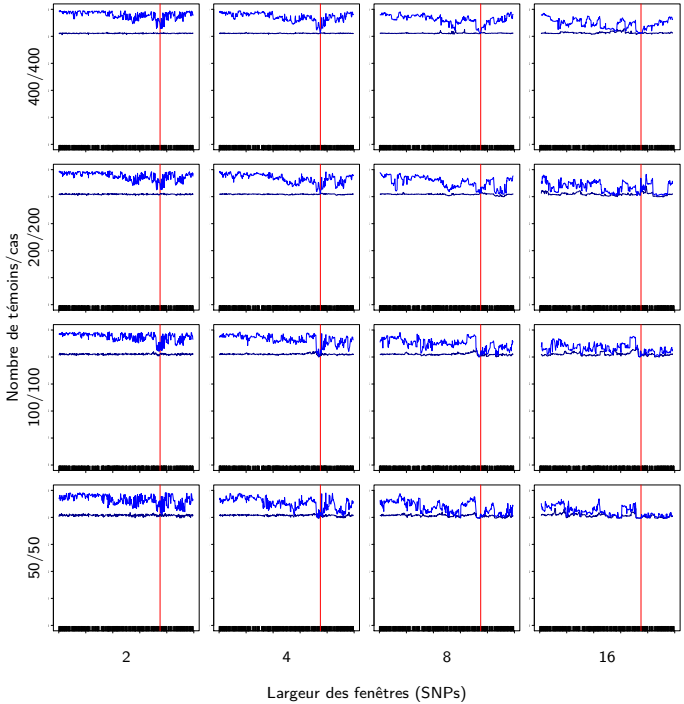


Figure C.III Taux de succès globaux en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **1,1**. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

⋘ < ⋙

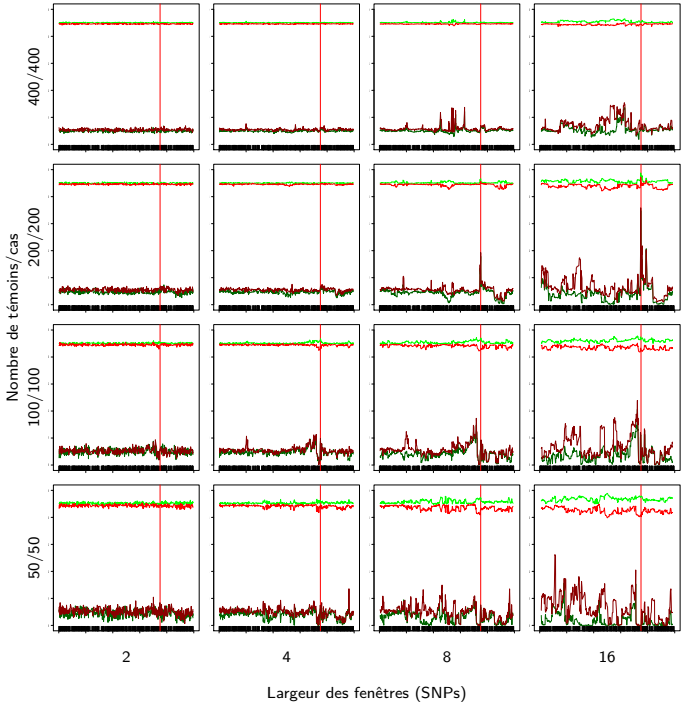


Figure C.IV Taux de succès partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **1,1**. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

⋘ < ⋙

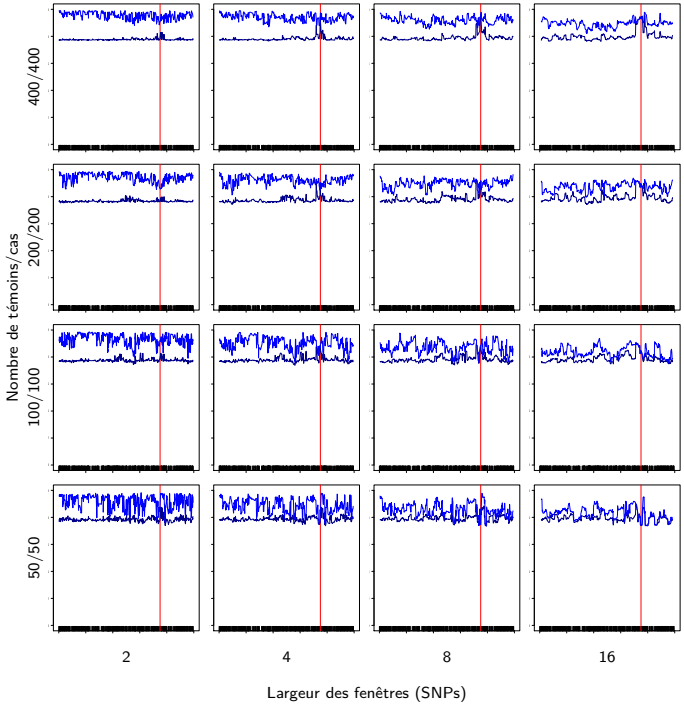


Figure C.V Taux de succès globaux en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **2**. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

⋘ < ⋙

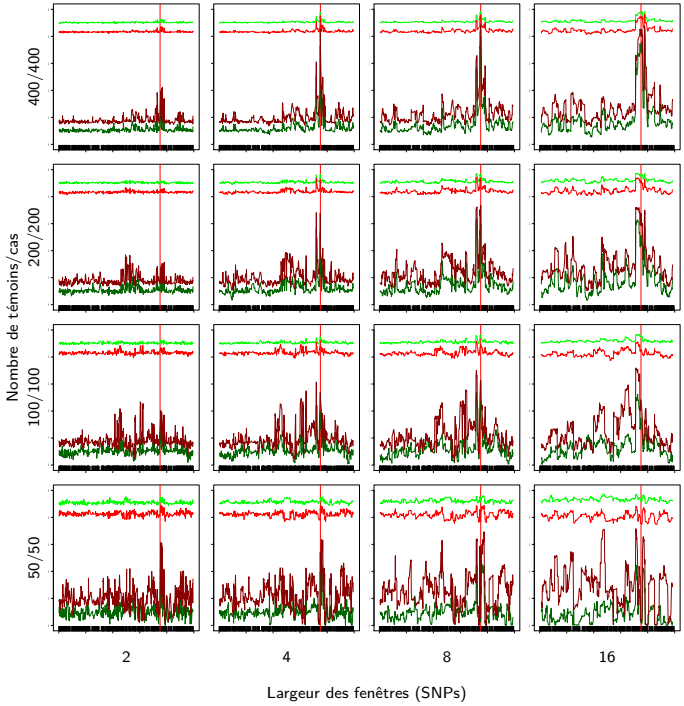


Figure C.VI Taux de succès partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **2**. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

≪ < ≫

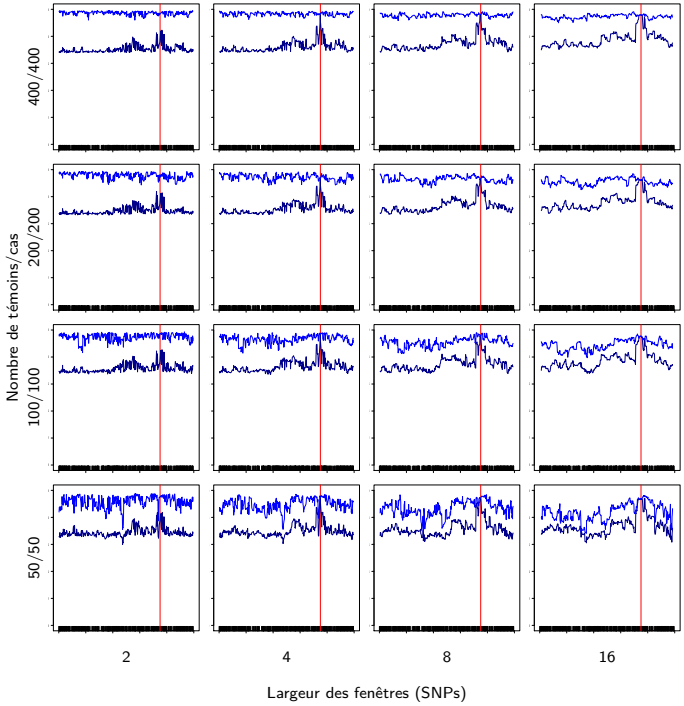


Figure C.VII Taux de succès globaux en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **10**. Échelle des ordonnées : $[0, 1]$. *Bleu :* Taux global utilitaire ($\pi_{\text{utilitaire}}$) ; *Bleu foncé :* Taux global (π).

≪ < ≫

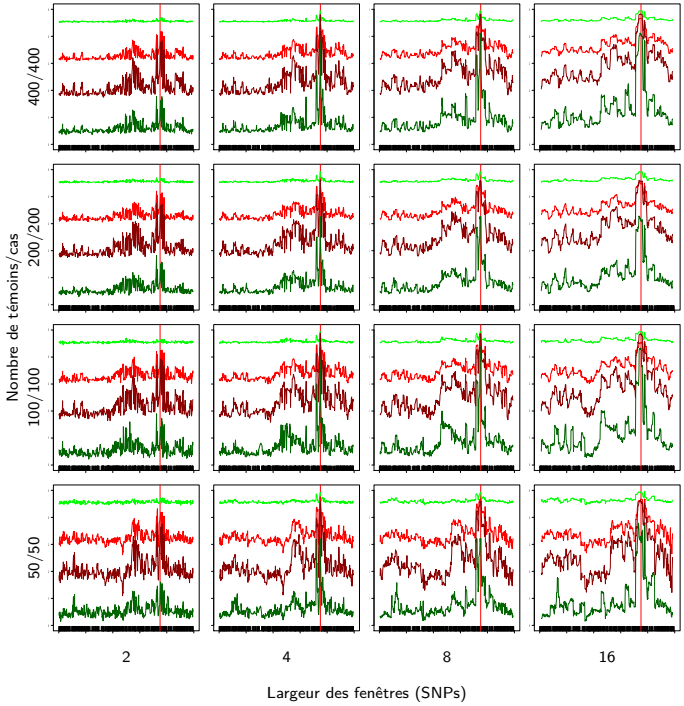


Figure C.VIII Taux de succès partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **10**. Échelle des ordonnées : $[0,\ 1]$. *Vert :* Témoins primitifs (π^0_{tem}) ; *Rouge :* Cas primitifs (π^0_{cas}) ; *Rouge foncé :* Cas mutants (π^1_{cas}) ; *Vert foncé :* Témoins mutants (π^1_{tem}).

≪ < ≫

APPENDICE D

TAUX DE SUCCÈS SEMI-PARTIELS, PAR TAILLE DE L'ÉCHANTILLON

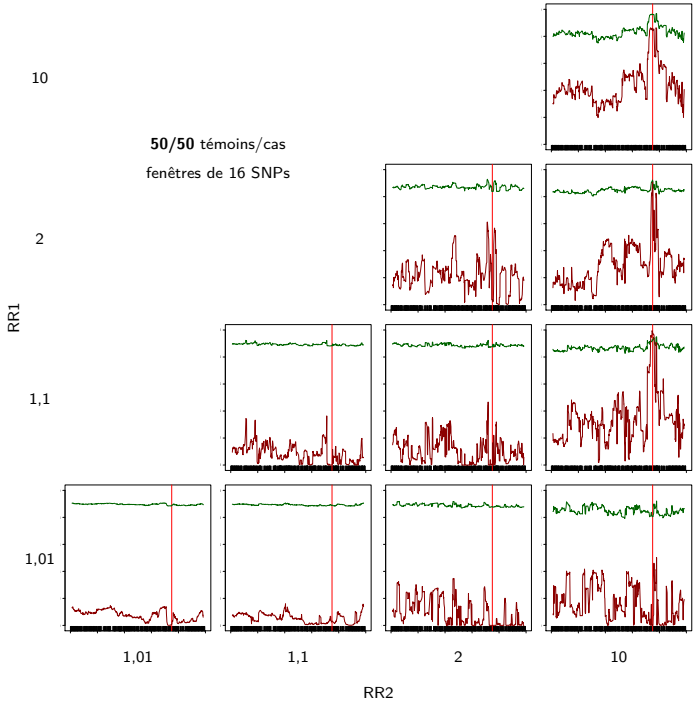


Figure D.I Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **50/50** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

⋘ < ⋙

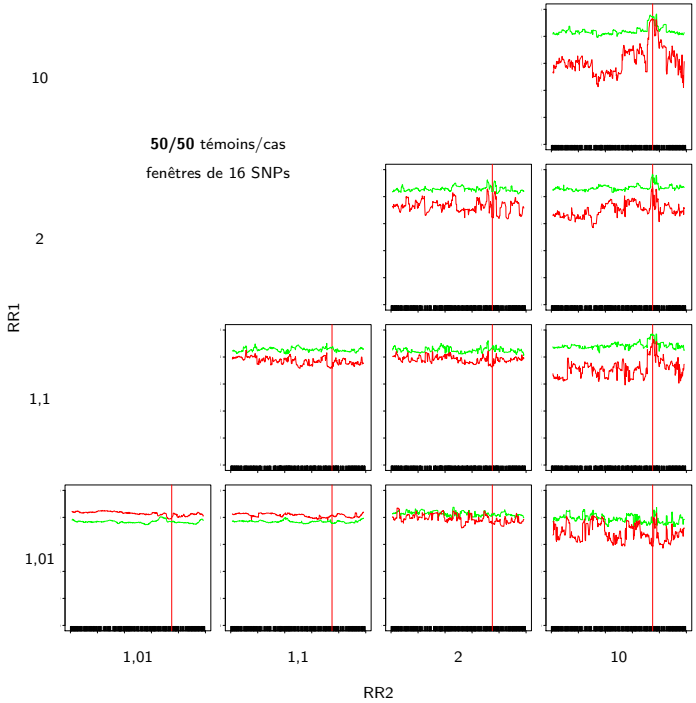


Figure D.II Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **50/50** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

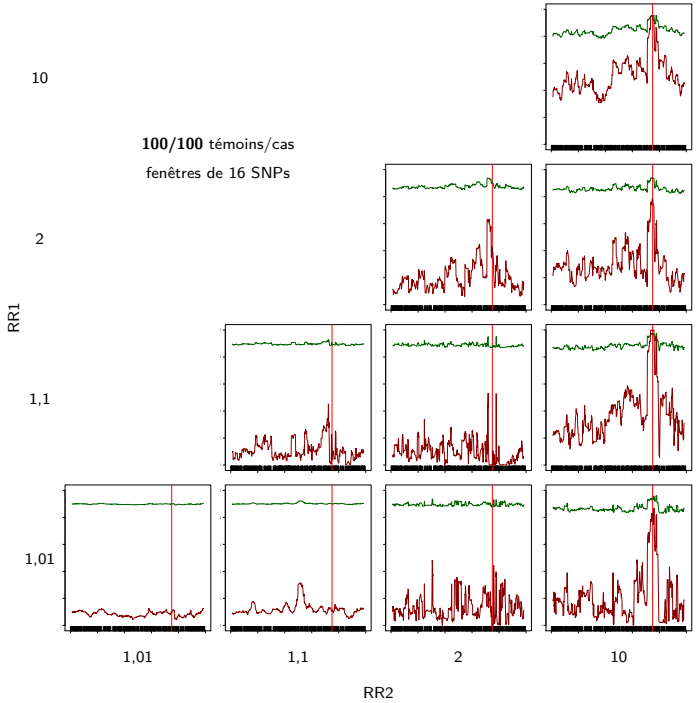


Figure D.III Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **100/100** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

⋘ < ⋙

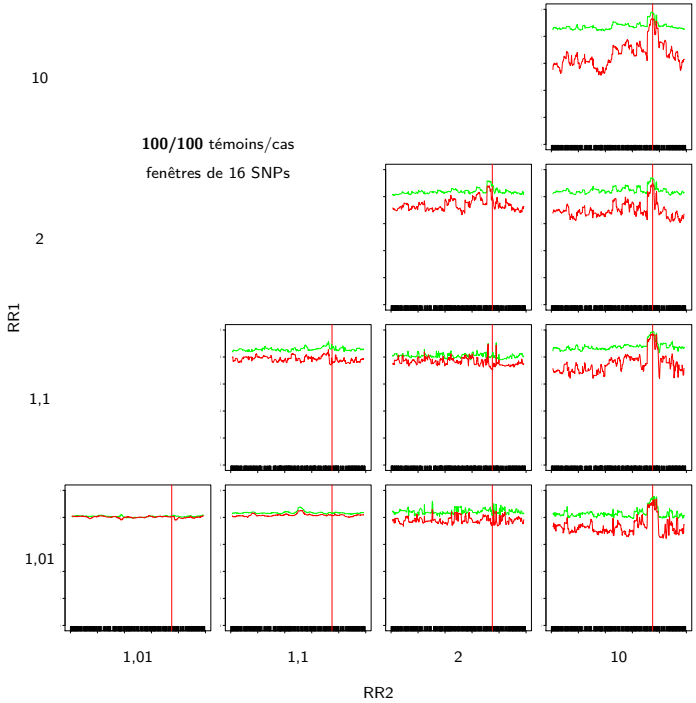


Figure D.IV Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **100/100** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

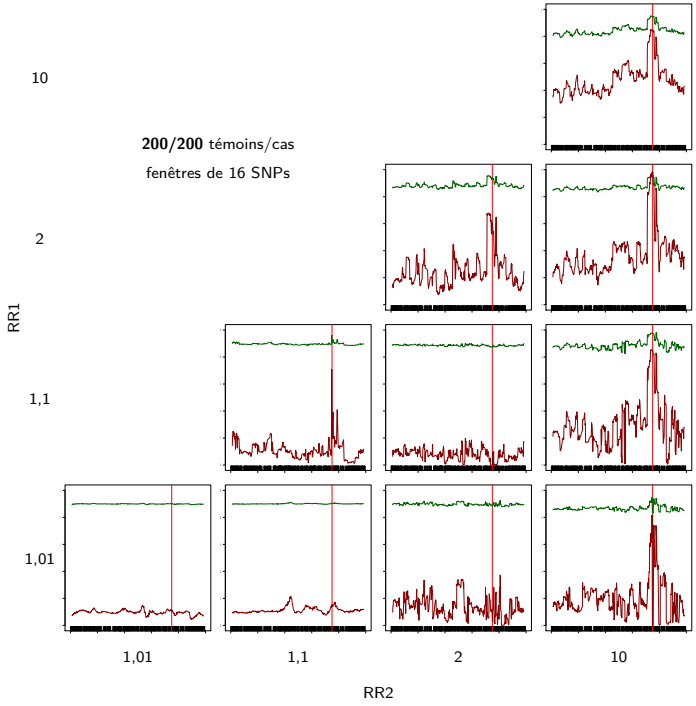


Figure D.V Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **200/200** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

⋘ < ⋙

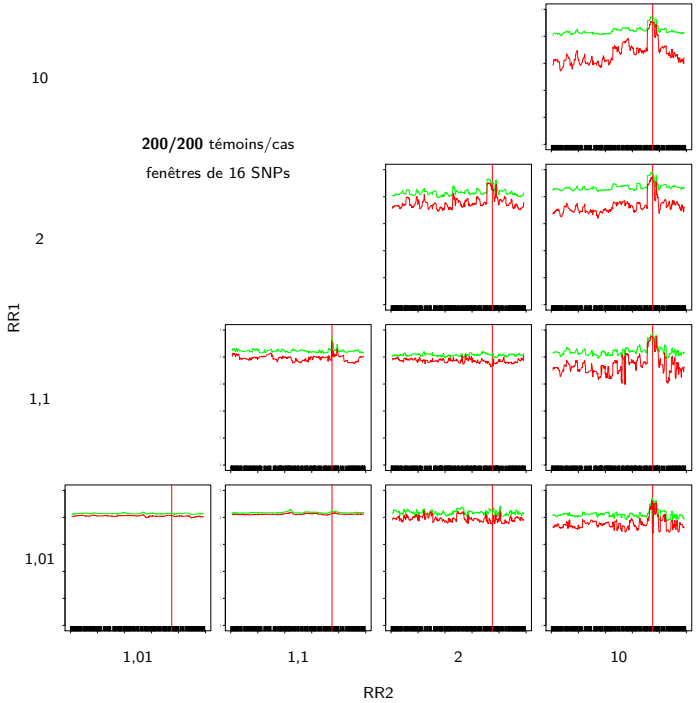


Figure D.VI Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **200/200** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

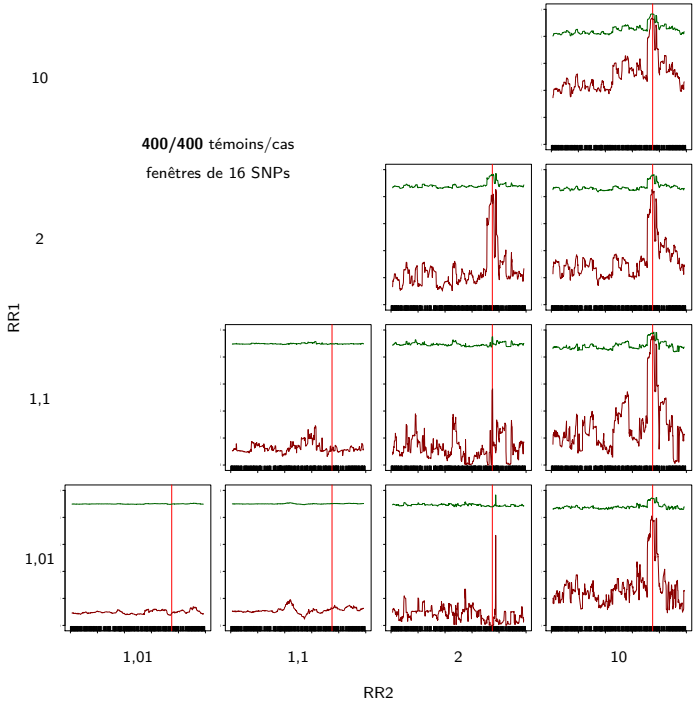


Figure D.VII Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **400/400** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

≪ < ≫

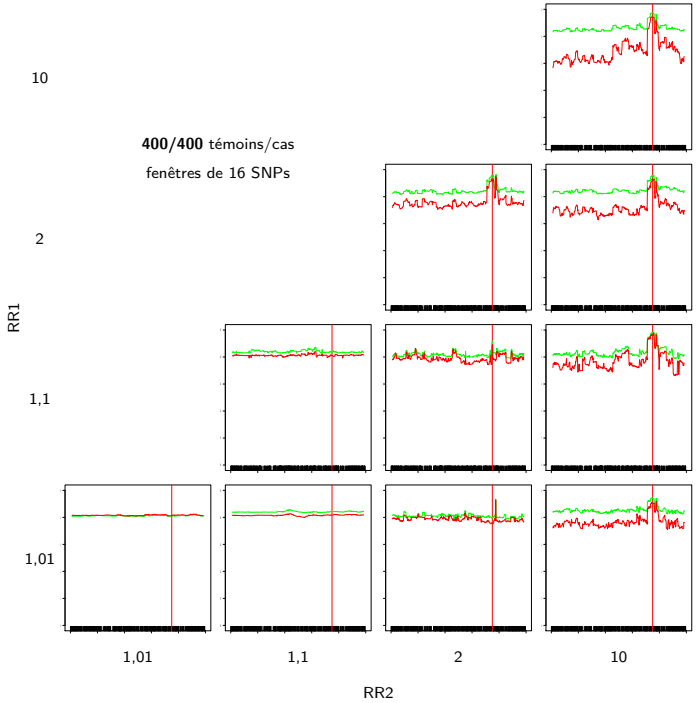


Figure D.VIII Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de **400/400** témoins/cas et des fenêtres de 16 SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

APPENDICE E

TAUX DE SUCCÈS SEMI-PARTIELS, PAR LARGEUR DES FENÊTRES

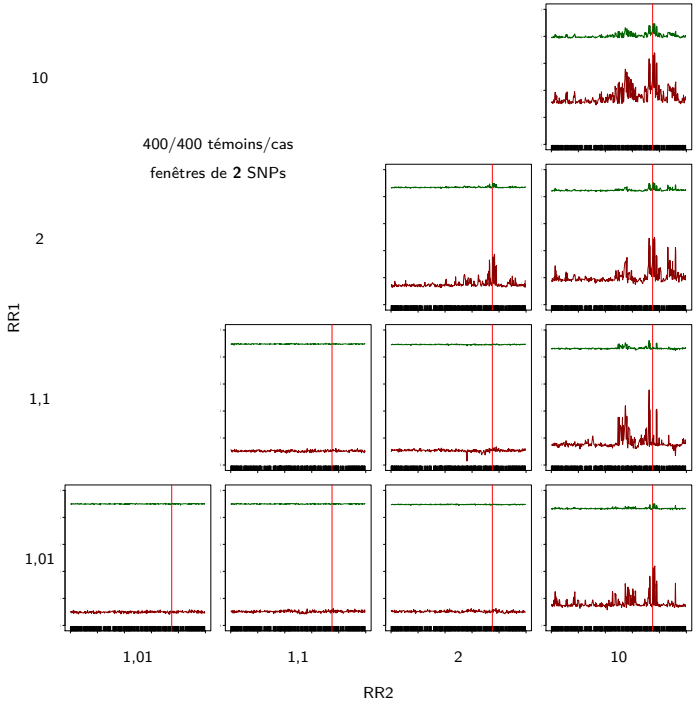


Figure E.I Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **2** SNPs. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

≪ < ≫

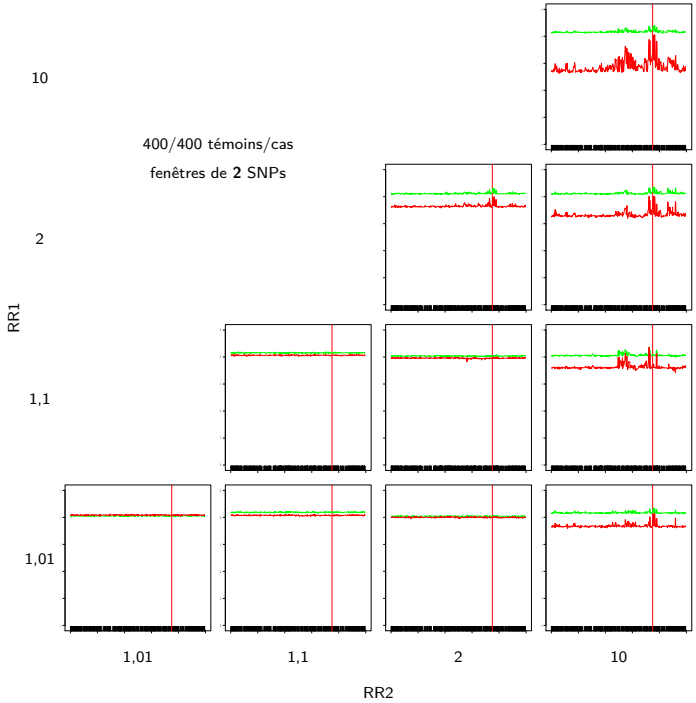


Figure E.II Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **2** SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

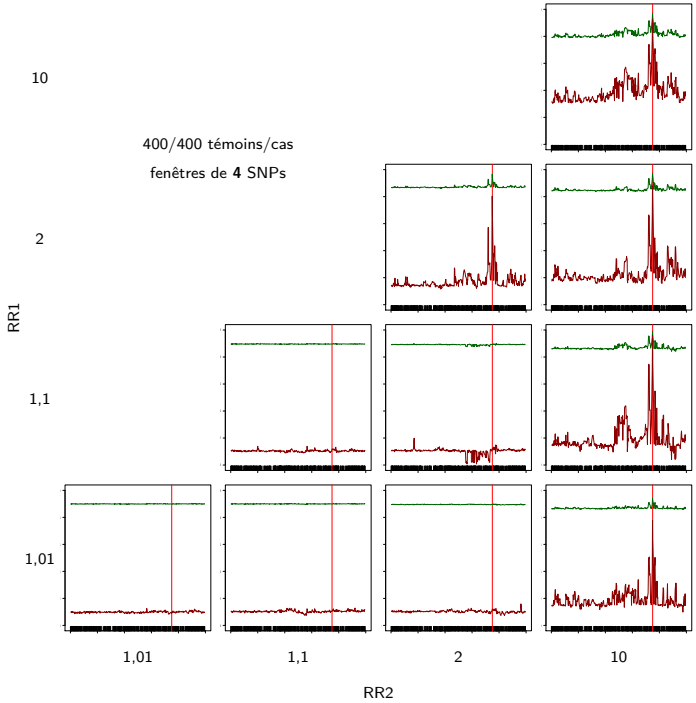


Figure E.III Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **4** SNPs. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

⋘ < ⋙

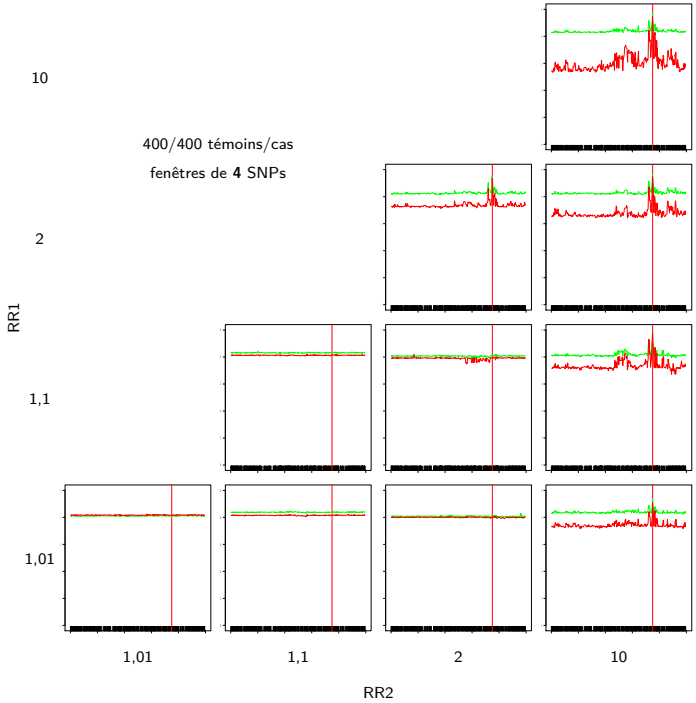


Figure E.IV Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **4** SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

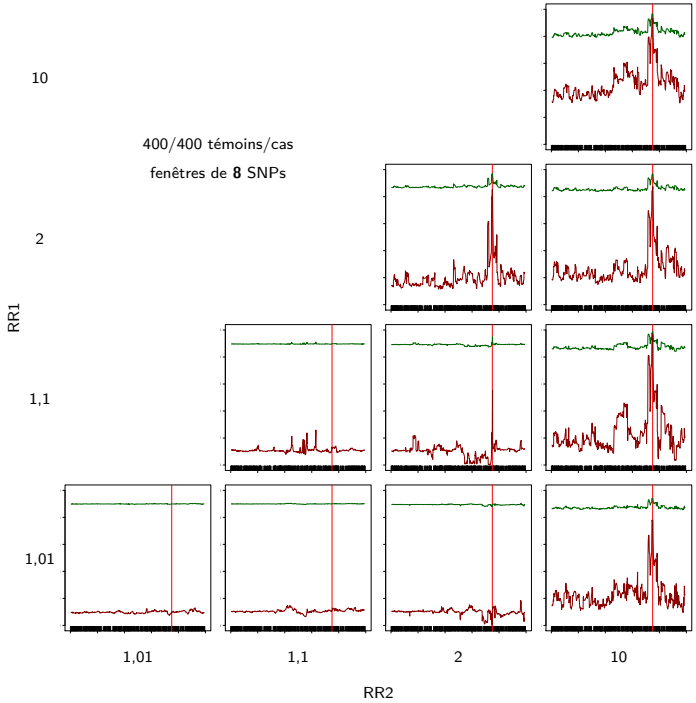


Figure E.V Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **8** SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

⋘ < ⋙

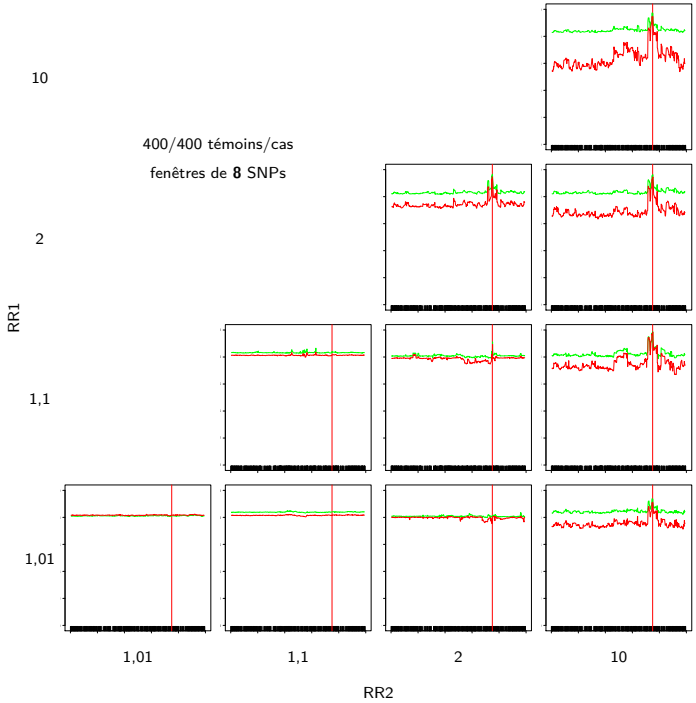


Figure E.VI Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **8** SNPs. Échelle des ordonnées : $[0,\ 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

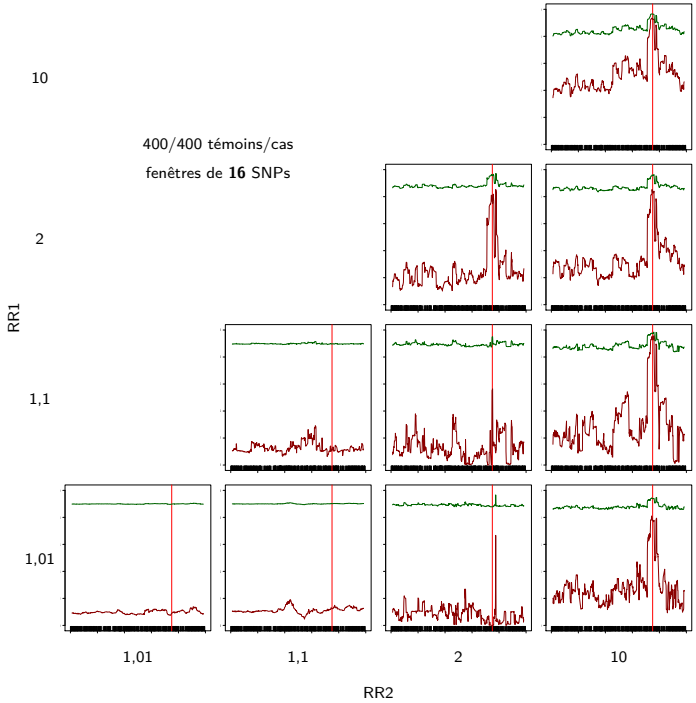


Figure E.VII Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **16** SNPs. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

⋘ < ⋙

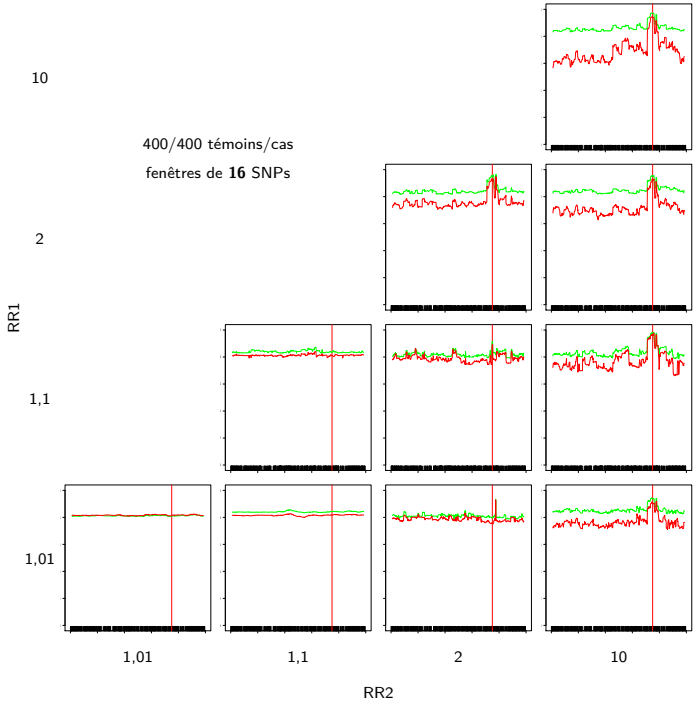


Figure E.VIII Taux de succès semi-partiels en fonction des risques relatifs RR1 et RR2, pour une taille de 400/400 témoins/cas et des fenêtres de **16** SNPs. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

≪ < ≫

APPENDICE F

TAUX DE SUCCÈS SEMI-PARTIELS, PAR RISQUES RELATIFS

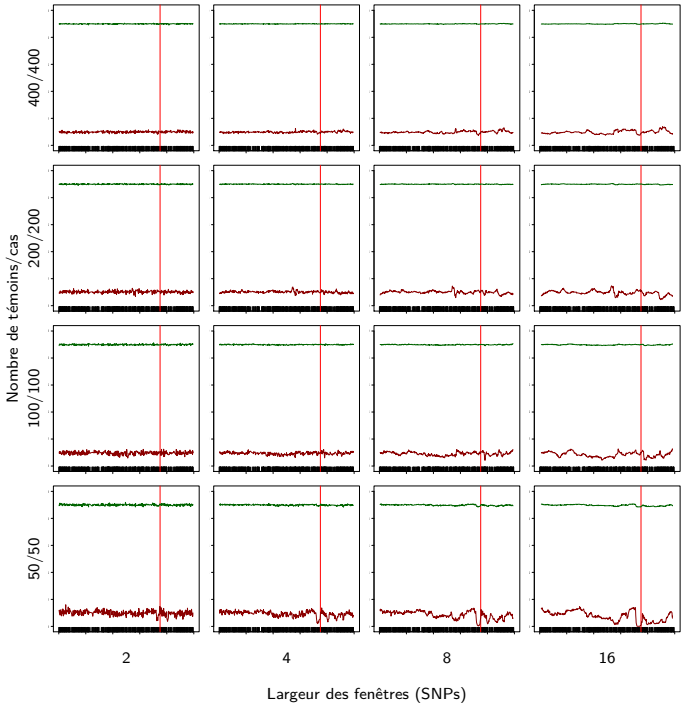


Figure F.I Taux de succès semi-partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **1,01**. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

⋘ < ⋙

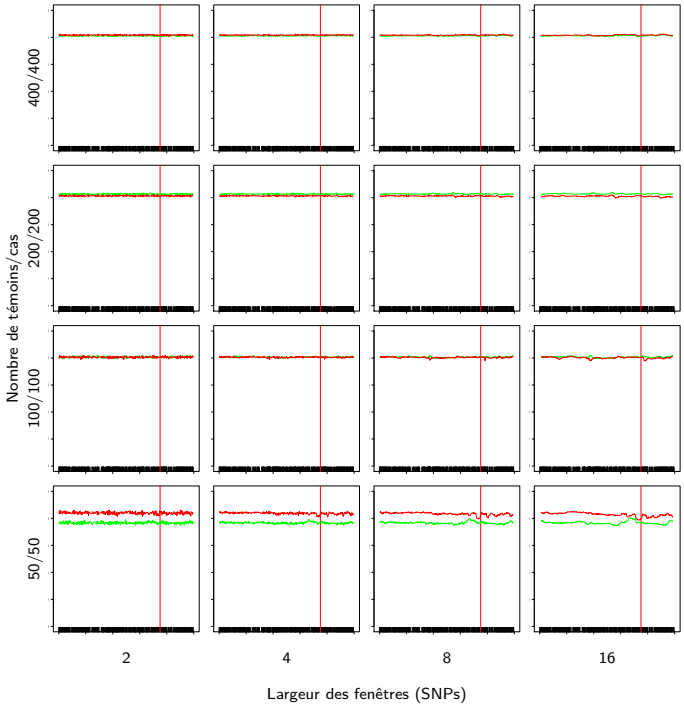


Figure F.II Taux de succès semi-partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **1,01**. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

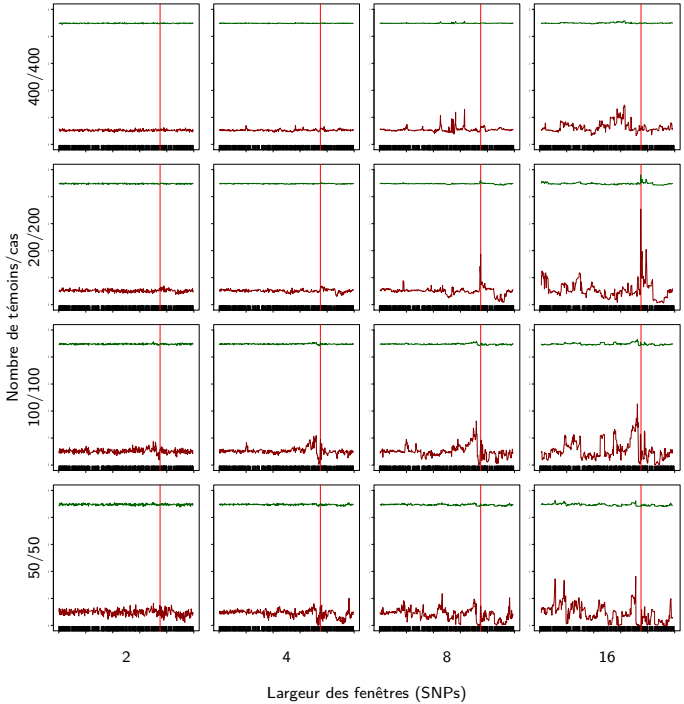


Figure F.III Taux de succès semi-partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **1,1**. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

≪ < ≫

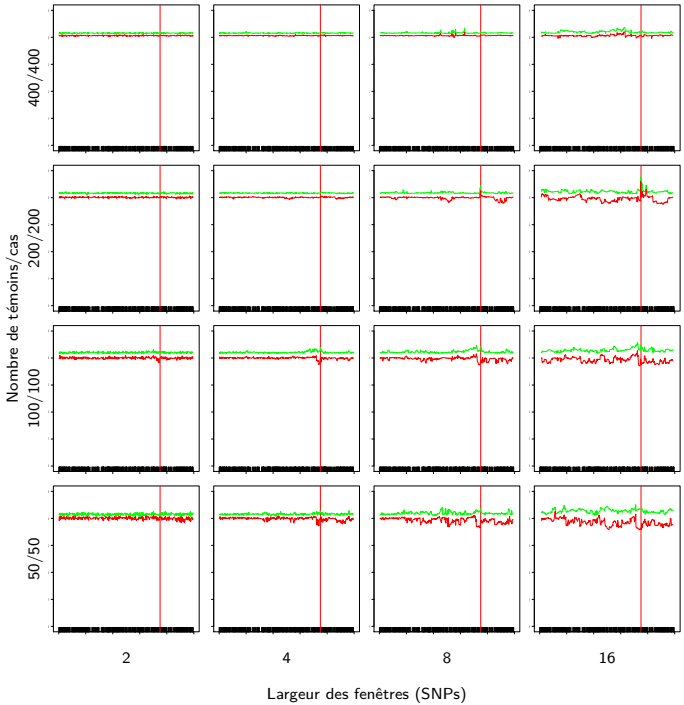


Figure F.IV Taux de succès semi-partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **1,1**. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

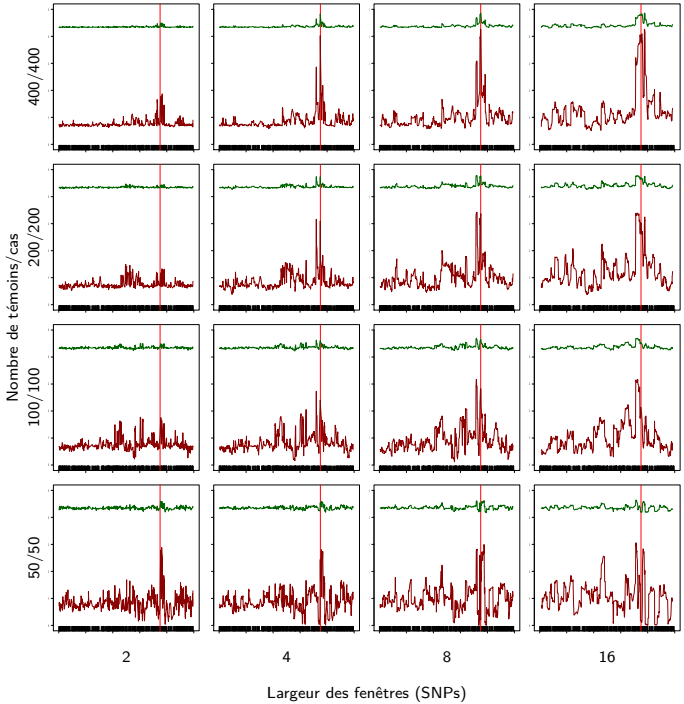


Figure F.V Taux de succès semi-partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **2**. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

⋘ < ⋙

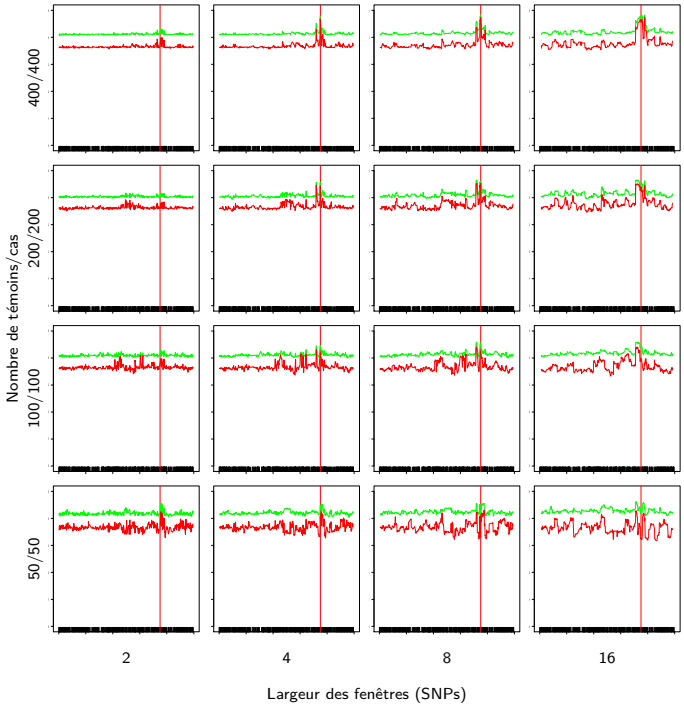


Figure F.VI Taux de succès semi-partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **2**. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

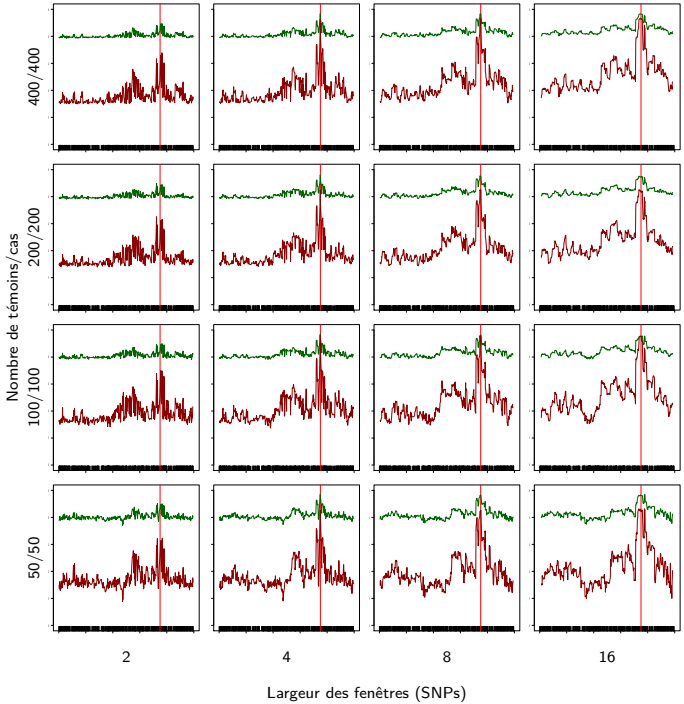


Figure F.VII Taux de succès semi-partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **10**. Échelle des ordonnées : $[0, 1]$. *Vert foncé :* Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1).

≪ < ≫

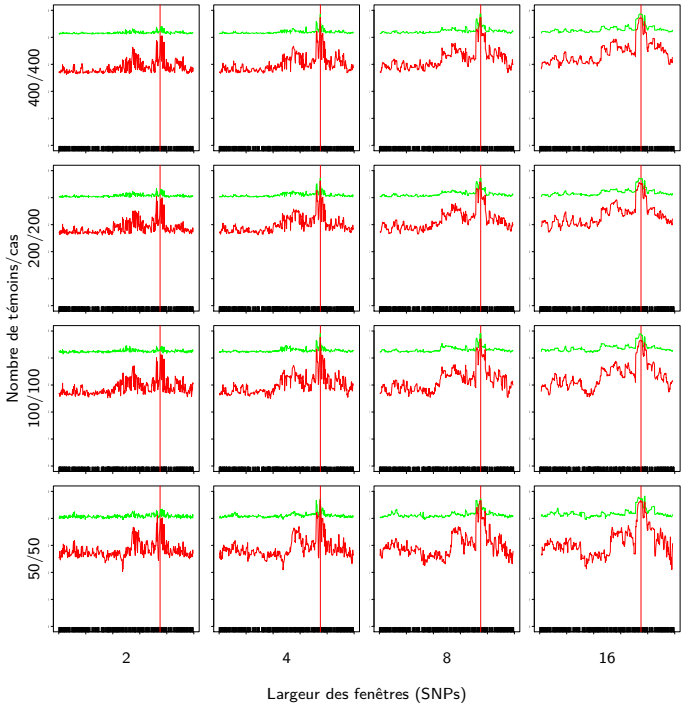


Figure F.VIII Taux de succès semi-partiels en fonction de la taille de l'échantillon et des fenêtres, pour RR1 = RR2 = **10**. Échelle des ordonnées : $[0, 1]$. *Vert :* Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

APPENDICE G

TAUX DE SUCCÈS SEMI-PARTIELS

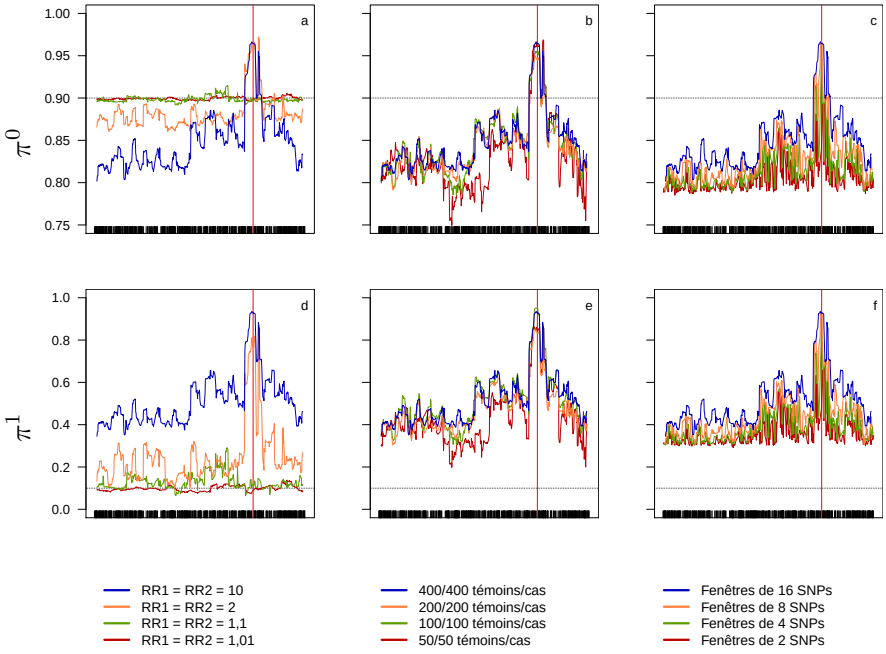


Figure G.I Taux de succès des primitifs (*a*,*b*,*c*) et des mutants (*d*,*e*,*f*) en fonction des risques relatifs RR1 et RR2 combinés (*a*,*d*), de la taille de l'échantillon (*b*,*e*) et de la largeur des fenêtres (*c*,*f*). Pour une rangée donnée, l'échelle des ordonnées est la même. La ligne pointillée représente le taux de succès aléatoire (*a*,*b*,*c* : 0,9 ; *d*,*e*,*f* : 0,1).
a,*d* : 400/400 témoins/cas, fenêtres de 16 SNPs ; *b*,*e* : RR1 = RR2 = 10, fenêtres de 16 SNPs ; *c*,*f* : RR1 = RR2 = 10, 400/400 témoins/cas.

≪ < ≫

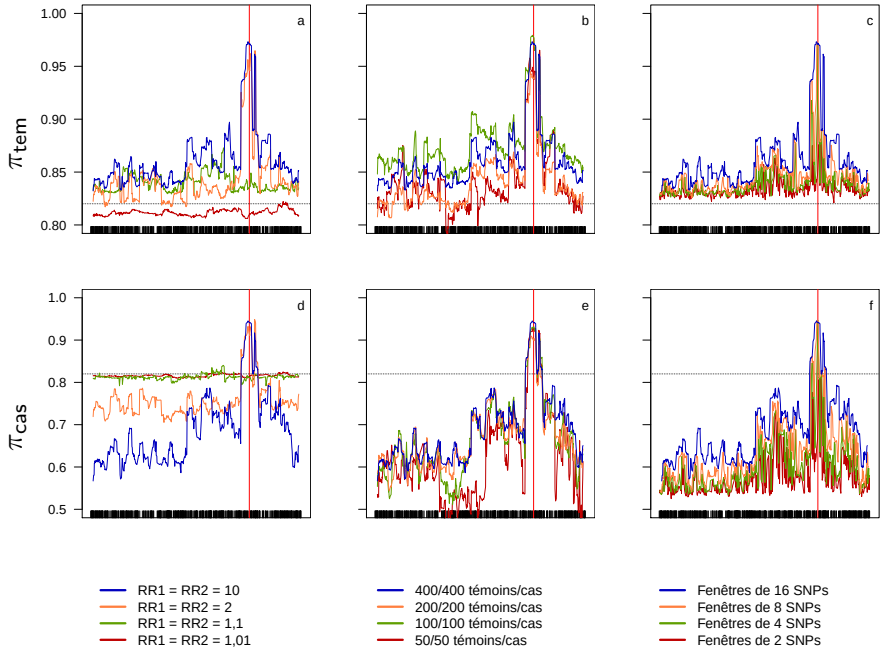


Figure G.II Taux de succès des témoins (*a*,*b*,*c*) et des cas (*d*,*e*,*f*) en fonction des risques relatifs RR1 et RR2 combinés (*a*,*d*), de la taille de l'échantillon (*b*,*e*) et de la largeur des fenêtres (*c*,*f*). Pour une rangée donnée, l'échelle des ordonnées est la même. La ligne pointillée représente le taux de succès aléatoire (0,82).
a,*d* : 400/400 témoins/cas, fenêtres de 16 SNPs ; *b*,*e* : RR1 = RR2 = 10, fenêtres de 16 SNPs ; *c*,*f* : RR1 = RR2 = 10, 400/400 témoins/cas.

APPENDICE H

TAUX DE SUCCÈS PÉRITIMS SEMI-PARTIELS

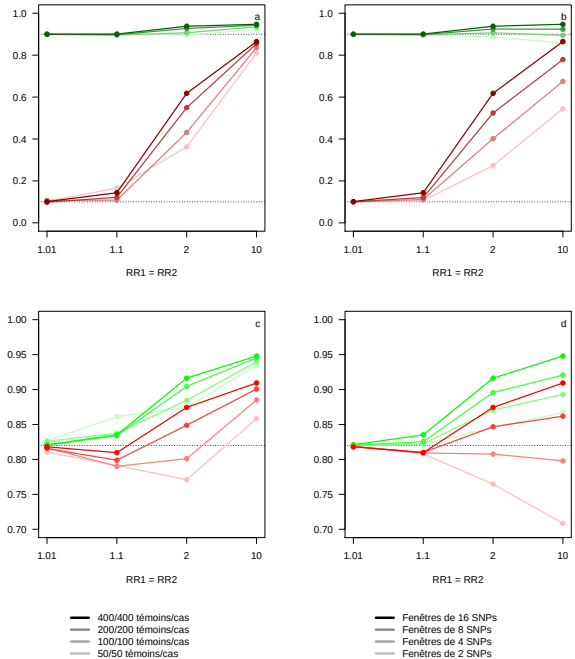


Figure H.I Taux de succès périTIMs semi-partiels en fonction des risques relatifs RR1 et RR2 combinés, par taille de l'échantillon (a,c) et par largeur de fenêtre (b,d). Chaque point représente la moyenne d'un taux de succès périTIM sur les 100 populations. Les lignes pointillées représentent les taux de succès aléatoires (a,b : 0,1 et 0,9 ; c,d : 0,82).
a,c : fenêtres de 16 SNPs ; b,d : 400/400 témoins/cas.
Vert foncé : Primitifs (π^0) ; *Rouge foncé :* Mutants (π^1) ;
Vert : Témoins (π_{tem}) ; *Rouge :* Cas (π_{cas}).

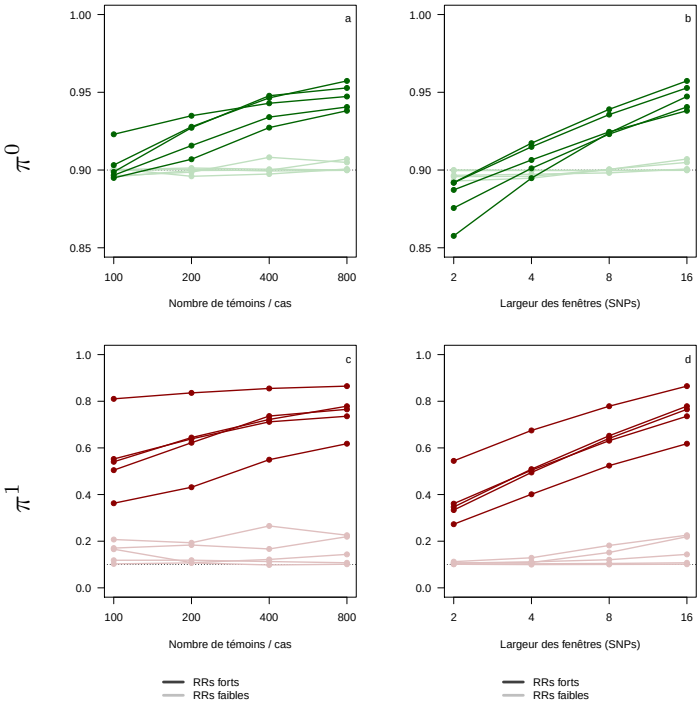


Figure H.II Taux de succès périTIMs des primitifs (*a*,*b*) et des mutants (*c*,*d*) en fonction de la taille de l'échantillon (*a*,*c*) et de la largeur des fenêtres (*b*,*d*), par risques relatifs. Chaque point représente la moyenne d'un taux de succès périTIM sur les 100 populations. La ligne pointillée représente le taux de succès aléatoire (*a*,*b* : 0,9 ; *c*,*d* : 0,1).
a,*c* : fenêtres de 16 SNPs ; *b*,*d* : 400/400 témoins/cas.

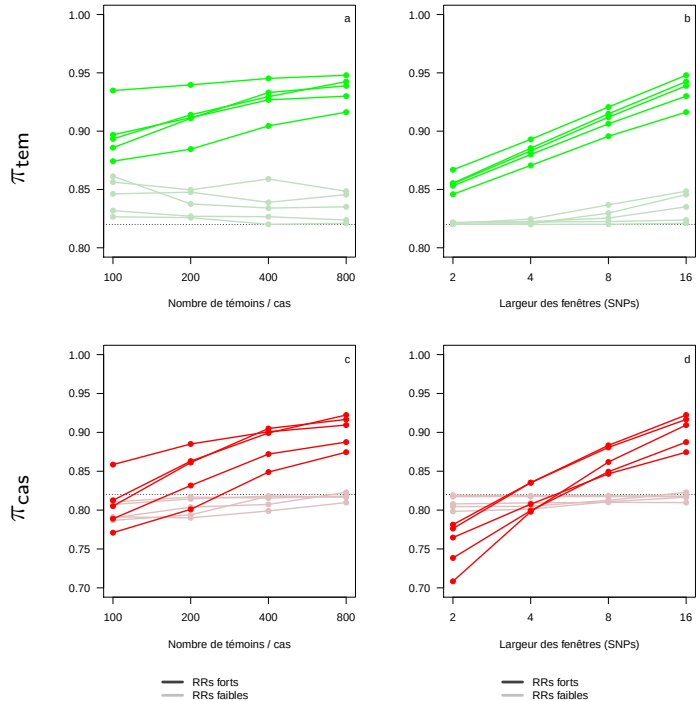


Figure H.III Taux de succès périTIMs des témoins (*a*,*b*) et des cas (*c*,*d*) en fonction de la taille de l'échantillon (*a*,*c*) et de la largeur des fenêtres (*b*,*d*), par risques relatifs. Chaque point représente la moyenne d'un taux de succès périTIM sur les 100 populations. La ligne pointillée représente le taux de succès aléatoire (0,82).
a,*c* : fenêtres de 16 SNPs ; *b*,*d* : 400/400 témoins/cas.

RÉFÉRENCES

Boucher, G. (2009). Intégration de la réalité diploïde et des modèles de pénétrance à une méthode de cartographie génétique fine. Mémoire de maîtrise, Université du Québec à Montréal.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Dawkins, R. (1976). *The selfish gene*. Oxford University Press, New York.

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. 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. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Griffiths, R. et Tavaré, S. (1994a). Ancestral inference in population genetics. *Statistical Science*, pages 307–319.

Griffiths, R. et Tavaré, S. (1994b). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1310):403–410.

Griffiths, R. et Tavaré, S. (1994c). Simulating probability distributions in the coalescent. *Theoretical Population Biology*, 46(2):131–159.

Griffiths, R. C. et Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3(4):479–502.

Hein, J., Schierup, M. H. et Wiuf, C. (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23(2):183–201.

Kingman, J. F. C. (1982). 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*, 156:1461–1463.

Kuhner, M., Yamato, J. et Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics*, 140(4):1421–1430.

Larribe, F. (2003). Cartographie génétique fine par le graphe de recombinaison ancestral. Thèse de doctorat, Université de Montréal.

Larribe, F. et Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, 21(1):43–69.

Larribe, F. et Lessard, S. (2008). A composite-conditional-likelihood approach for gene mapping based on linkage disequilibrium in windows of marker loci. *Stat Appl Genet Mol Biol*, 7(1):Article27.

Larribe, F., Lessard, S. et Schork, N. J. (2002). Gene mapping via the ancestral recombination graph. *Theor Popul Biol*, 62(2):215–229.

Mendel, J. G. (1865). Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, 4:3–47.

Stephens, M. et Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43–59.

Vahey, S. (2008). Modélisation des paramètres de pénétrance incomplète et de phénocopie d'une méthode de cartographie fine d'une maladie complexe. Mémoire de maîtrise, Université du Québec à Montréal.

Wakeley, J. (2009). *Coalescent theory: an introduction*. Roberts & Co. Publishers.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16:97–159.