UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Intégration de la réalité diploïde et des modèles de pénétrance à une méthode de cartographie génétique fine

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

GABRIELLE BOUCHER

AOÛT 2009

REMERCIEMENTS

Je remercie Fabrice Larribe, mon directeur de recherche, pour sa disponibilité, son soutien et sa confiance. Merci de m'avoir proposé un défi aussi captivant. La passion que tu mets à ton travail est inspirante et contagieuse, attention à l'épidémie. Merci pour chacune des heures que tu as passées à déchiffrer mes calculs, démêler mes idées et corriger ce mémoire. Je persiste tout de même à croire que la nuit n'est pas faite pour travailler...

Je remercie les professeurs du département de l'UQAM pour leur dévouement et leur compétence. Vous méritez l'appui et le support de vos étudiants. Un merci particulier à Sorana Froda, qui m'a encouragée plus que fortement à entreprendre une maîtrise. Vous aviez raison, évidemment. Merci aussi à René Ferland et Alain Desgagné, membres du jury, pour leurs corrections et commentaires qui m'ont permis d'améliorer la version finale de ce mémoire.

Merci à ma famille et mes amis, qui placent en moi une confiance aveugle et inexplicable, mais appréciée. Un merci particulier à mes parents. Vous avez raison d'être fiers, puisque vous faites partie de tous mes succès.

Merci à Sylvain, mon amour et mon soleil. Dix ans ont passé, et ton sourire me fait toujours craquer. Merci pour ta compréhension, même lorsque je suis incompréhensible.

Merci à Raphaël, mon rayon de bonheur. Merci pour ton innocence et ta joie de vivre. Ma petite réussite en génétique appliquée...

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Nous présentons dans ce mémoire des outils permettant de généraliser une méthode de cartographie génétique fine. Nous y résumons les concepts de base de la statistique génétique et y décrivons aussi la méthode de cartographie génétique fine que nous cherchons à généraliser en permettant l'utilisation de génotypes plutôt que d'haplotypes. Pour ce faire, nous comparons diverses méthodes reconnues d'estimation d'haplotypes. Le développement nouveau de ce travail consiste en un algorithme EM conditionnel aux phénotypes permettant d'estimer les haplotypes associés à un échantillon de génotype, ainsi que le statut au gène causal du caractère étudié. Nous généralisons la méthode de cartographie par l'ajout d'étapes au modèle d'échantillonnage pondéré. Nous effectuons finalement quelques tests par simulation.

MOTS-CLÉS : algorithme EM, cartographie génétique, coalescence, diplotype, échantillonnage pondéré, estimation, génotype, gène causal, haplotype, modèle de pénétrance, phénotype, vraisemblance composite

INTRODUCTION

Depuis plusieurs années, on note un intérêt marqué pour l'étude de la grammaire de la vie, la génétique. Un des objectifs de cette vaste discipline est d'établir une carte des gènes impliqués dans la transmission des caractères héréditaires. L'évolution des connaissances et des outils technologiques dans ce domaine va de pair avec l'élaboration d'outils statistiques permettant d'analyser les données recueillies en laboratoire. Ainsi, plusieurs méthodes statistiques de cartographie génétique ont été développées. Certaines d'entre elles reposent toutefois sur des hypothèses peu réalistes quant à la forme que prennent les échantillons ou les modèles biologiques impliqués.

Ce mémoire a pour objectif de généraliser une méthode de cartographie génétique fine, MapArg (Larribe, Lessard et Schork, 2002), afin de permettre l'intégration de la réalité diploïde et des modèles de pénétrance. Cette généralisation est importante, puisqu'elle permettrait éventuellement l'application de la méthode à des échantillons provenant de populations humaines.

Dans le premier chapitre, nous présenterons les concepts de base de la statistique génétique, de manière à familiariser le lecteur avec les termes et modèles nécessaires à la compréhension du sujet traité. Le second chapitre portera quant à lui sur la méthode de cartographie que nous cherchons à généraliser. Nous résumerons au troisième chapitre les principales méthodes d'estimation des haplotypes, en présentant les avantages et limites de chacune. Au quatrième chapitre, nous développerons une nouvelle méthode d'estimation, conditionnelle aux phénotypes. Le cinquième chapitre mettra les pièces en place, en intégrant la réalité diploïde à la méthode de cartographie. Finalement, nous présenterons au sixième chapitre quelques résultats de simulations.

CHAPITRE I

CONCEPTS DE BASE

La statistique génétique comporte deux volets. Le premier, la génétique, comprend un ensemble de concepts biologiques et chimiques impliqués dans la transmission des caractères héréditaires. Le second, la statistique, consiste en une modélisation mathématique de la réalité biologique dans le but de résoudre les problèmes posés. Ce chapitre se veut une présentation concise des concepts de biologie, de génétique et de mathématiques nécessaires à la compréhension de la problématique traitée et des solutions apportées. La plupart des informations de nature biologique qui y sont présentées ont été tirées du livre *Biologie* de Neil A. Campbell et Richard Mathieu (1995) et des notes de Almgren *et al.* (2003).

1.1 Introduction à la génétique

1.1.1 Hérédité, gènes et ADN

L'hérédité a été remarquée depuis des siècles, sinon des millénaires. C'est elle qui explique la ressemblance, parfois flagrante, entre les parents et leurs enfants. Parfois, elle est responsable de prédispositions familiales à certaines maladies. De façon plus précise, le concept d'hérédité fait référence à l'ensemble des caractères transmis par les parents à leurs enfants au moment de la reproduction. Bien que le principe de l'hérédité soit connu depuis longtemps, la science n'en a découvert les fondements que récemment, par l'étude de la génétique.

Depuis les travaux de Gregor Mendel au XIX^e siècle (Mendel, 1866), on reconnaît que les caractères héréditaires sont transmis par des unités, les gènes, qui demeurent distinctes de générations en générations et dont l'expression détermine le phénotype de l'individu. Ces gènes se présentent sous différentes formes, les allèles, pouvant donner lieu à des variations dans les caractères héréditaires. Chaque être humain possède deux allèles pour chaque gène, hérités du père et de la mère. Tout comme Mendel, nous étudierons le cas de phénotypes dichotomiques, c'est-à-dire se présentant sous deux formes distinctes, chacune associée à un allèle. Un caractère de ce type peut être récessif, c'est-à-dire qu'il nécessite deux copies de l'allèle causal pour s'exprimer, ou dominant, si la présence d'une seule copie suffit. Il existe aussi des caractères codominants, pour lesquels un phénotype intermédiaire peut être observé lorsqu'une unique copie du gène causal est présente. L'expression des gènes associés à un caractère peut aussi être déterminée par un modèle probabiliste, le modèle de pénétrance. Ce modèle comporte trois paramètres : f_0, f_1 et f_2, respectivement la probabilité qu'un individu présente le phénotype d'intérêt s'il porte aucune, une seule ou deux copies de l'allèle causal. Un individu qui présente un caractère normalement associé à un allèle sans en être porteur constitue une phénocopie.

Les gènes, porteurs de l'hérédité, ont eux-mêmes un support physique, la molécule d'ADN (acide désoxyribonucléique). L'ADN est une molécule contenue dans chaque cellule et se présentant sous la forme d'une longue chaîne de nucléotides. Chaque nucléotide est identifié par la base azotée qu'il contient : l'adénine (A), la thymine (T), la guanine (G) ou la cytosine (C). La combinaison successive et linéaire de ces quatre bases détermine l'ensemble du matériel génétique d'un individu. Ainsi, chaque gène correspond à une séquence précise de ce code à quatre lettres qu'est l'ADN.

Dans les cellules, l'ADN se répartit en un nombre variable de chromosomes selon les espèces. Ces chromosomes peuvent être isolés ou appariés. Ainsi, l'être humain possède 23 paires de chromosomes. Comme la plupart des plantes et des animaux, nous sommes des êtres diploïdes, c'est-à-dire que nos chromosomes se présentent en deux exemplaires dits homologues, puisqu'ils codent pour les mêmes caractères. Un des exemplaires provient du père, et l'autre, de la mère. Par opposition, les bactéries sont habituellement

haploïdes, c'est-à-dire qu'elles n'ont qu'un seul jeu de chromosomes. Certaines plantes et, exceptionnellement, des animaux, peuvent présenter une polyploïdie, c'est-à-dire plus de deux jeux de chromosomes. C'est le cas de la banane commerciale, qui est triploïde. Notons toutefois la particularité des chromosomes sexuels. Chez l'humain, il existe deux types de chromosomes sexuels, le X et le Y. Alors que la femme possède deux exemplaires du chromosome X, l'homme n'en possède qu'un seul, qui est jumelé à un chromosome Y.

1.1.2 Recombinaison et mutation

La diversité biologique observable implique des mécanismes de modification du code génétique des individus. Un de ces mécanismes est le partage aléatoire des chromosomes homologues. Au moment de la reproduction, chacun des parents transmet à son enfant un seul jeu de chromosomes. Le partage aléatoire des chromosomes homologues en deux jeux distincts se fait au moment de la méïose. La méïose est une division cellulaire particulière aux cellules reproductrices, les gamètes, par laquelle les chromosomes homologues sont copiés et séparés. Puisque le partage des paires se fait de manière aléatoire, ce brassage du génome permet à lui seul 2^{23} associations différentes, soit plus de huit millions de possibilités, pour un seul parent.

Un autre phénomène survenant au moment de la méiose, l'enjambement, ajoute un degré supplémentaire de brassage génétique. Les chromosomes homologues sont alors impliqués dans un transfert réciproque de matériel génétique. À la manière de chaînes, ceux-ci se croisent puis se «brisent» au même endroit, pour se reformer en échangeant les segments impliqués. On dit alors que les gènes situés de part et d'autre de l'enjambement ont subit une recombinaison, puisque la combinaison des allèles sur ceux-ci est modifiée. La figure 1.1 représente un enjambement de façon imagée, par analogie avec des chaînes. Plusieurs enjambements peuvent avoir lieu en même temps. En moyenne, il s'en produit environ deux ou trois par paire de chromosomes. Lorsqu'un nombre pair d'enjambements survient entre deux gènes, on n'observe pas de recombinaison entre ceux-ci, puisqu'ils retrouvent leur agencement initial. La figure 1.2 illustre un double enjambement. Notons

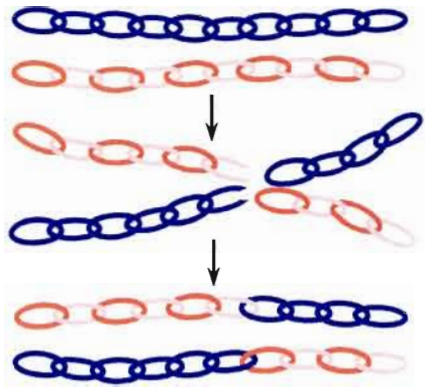

Figure 1.1 Illustration imagée d'un enjambement. Les deux chromosomes homologues sont représentés par des chaînes de couleurs différentes. Les gènes situés de part et d'autre de l'enjambement ont recombiné.

toutefois qu'il est rare que deux enjambements se produisent de manière très rapprochée. Le partage aléatoire et l'enjambement produisent un nombre pratiquement infini de combinaisons des gènes de chaque parent. Ainsi, il est virtuellement impossible que deux enfants de mêmes parents aient reçu exactement le même bagage génétique, exception faite des jumeaux identiques.

Nous venons d'expliquer comment la recombinaison et le partage aléatoire des chromosomes homologues sont responsables du brassage génétique à l'origine des différences entre frères et soeurs. Toutefois, ce brassage ne crée pas de nouvelles variantes de l'ADN. Celles-ci sont plutôt causées par la mutation. Une mutation est une erreur qui survient lors de la copie de la séquence d'ADN. Ainsi, une ou plusieurs bases peuvent être modifiées ou insérées dans le code original, donnant lieu à une séquence nouvelle différente de celle des parents. Il existe différents types de mutations, certains étant très rares, tandis que d'autres surviennent plus fréquemment. Certaines mutations peuvent mener à un phénotype nouveau ou à l'apparition d'un dysfonctionnement. Une mutation qui n'affecte pas la capacité d'un individu à se reproduire est dite neutre, en ce sens que sa transmission est entièrement régie par les lois du hasard. Généralement, ces mutations n'affectent pas les caractères héréditaires. Les événements de mutations étant considérés

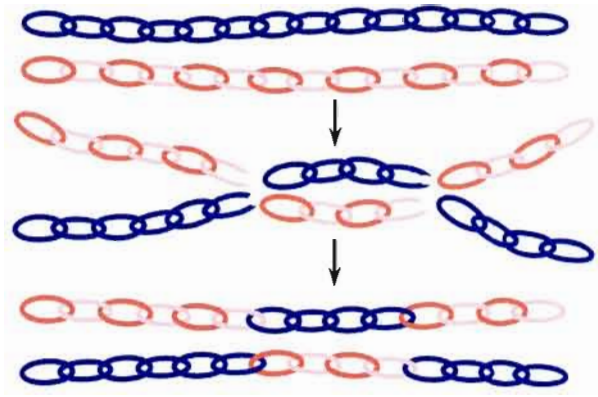

Figure 1.2 Illustration imagée d'un double enjambement. Les gènes situés aux extrémités n'ont pas recombiné.

plutôt rares, on juge habituellement improbable qu'une même mutation survienne chez plusieurs individus dans une courte période de temps.

1.2 Les balises du génome

1.2.1 Les marqueurs

Le code génétique humain comprend de longues séquences communes. Or, la génétique s'intéresse plus particulièrement à la variation entre les individus. Par conséquent, il est rarement nécessaire de conserver l'information génétique dans son ensemble. Pour cette raison, on fait généralement appel à des marqueurs génétiques. Ainsi, les marqueurs de type SNP (*single nucleotide polymorphism*) correspondent aux sites de variation d'une seule base du code génétique. Ce type de marqueur prend généralement une forme binaire. En effet, bien que l'alphabet de l'ADN comporte quatre lettres, les mutations n'impliquant qu'une seule base ne se produisent habituellement qu'une fois. Les marqueurs de type microsatellite sont quant à eux de courtes séquences d'ADN formées de la répétition de quelques bases. Un même marqueur de type microsatellite peut prendre plusieurs formes, selon le nombre de répétitions. D'autres types de marqueurs peuvent aussi être utilisés. Ainsi, on peut englober les gènes dans le concept de marqueurs,

puisque ceux-ci sont souvent synonymes de variations entre les individus. Les variantes du code génétique à un même marqueur sont nommées allèles et un individu est dit hétérozygote à un marqueur (autrement homozygote) si les deux allèles qu'il possède sur celui-ci sont différents.

Les allèles que porte un individu pour un ensemble donné de marqueurs forment son génotype. Chez les individus diploïdes, ce génotype est composé de deux haplotypes correspondant chacun à la séquence des allèles présents sur l'un des deux chromosomes homologues. De manière générale, les analyses en laboratoire permettent de déterminer le génotype, mais pas les deux haplotypes qui le composent. L'ensemble de l'information génétique étant combinée au moment de l'analyse, on peut déterminer quels sont les allèles d'un individu à chaque marqueur, mais pas de quelle manière ils sont agencés sur les deux chromosomes. Cet agencement, ou phase, formé des deux haplotypes, sera nommé diplotype, selon une notation récemment adoptée par la littérature. La figure 1.3 illustre ces concepts de façon schématique. On y représente les haplotypes comme des mots, le diplotype étant le couple formé des deux mots distincts. On remarque que le génotype comprend les mêmes lettres sur chaque ligne, mais que les deux mots n'y sont plus lisibles. De la même manière, les génotypes obtenus en laboratoire ne permettent habituellement pas de retrouver directement les haplotypes. Ce problème particulier est traité dans les chapitres 3 et 4.

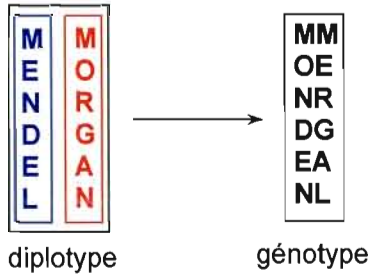

Figure 1.3 Représentation schématique d'un diplotype. On a ici 6 marqueurs, dont les lettres sont les allèles. Les deux mots représentent les haplotypes formant le diplotype. Combinées ensemble, les lettres forment le génotype.

1.2.2 Distances et recombinaisons

Nous avons vu précédemment que le partage des chromosomes homologues se faisait de manière aléatoire. Ainsi, deux marqueurs situés sur des paires de chromosomes différentes sont transmis de manière tout à fait indépendente. En fait, il y a alors 50% des chances pour que les allèles à ce marqueur soient transmis ensemble, ce qui correspond à choisir aléatoirement le même grand-parent pour les deux chromosomes légués à l'enfant. De même, il est probable que deux marqueurs situés sur le même chromosome soient plutôt transmis conjointement. De ce fait, lorsque deux marqueurs sont positionnés sur le même chromosome, on dit qu'ils sont liés. Cette liaison peut être plus ou moins forte selon la position relative des marqueurs. La position d'un marqueur dans le code génétique est nommée *locus* (*loci*, au pluriel). Plus les *loci* sont rapprochés, moins il est probable que les marqueurs soient divisés par une recombinaison.

La position relative de deux marqueurs peut être exprimée en morgans (M) ou, plus fréquemment, en centimorgans (cM), en hommage à Thomas Hunt Morgan. Au début du XX^e siècle, celui-ci posa les bases de la cartographie génétique en associant pour la première fois un gène à un chromosome. Les premières cartes du génome basées sur la recombinaison furent par la suite établies par l'un de ses étudiants, Alfred H. Sturtevant. Aujourd'hui encore, il est fréquent que l'on mesure les distances génétiques en terme de taux de recombinaison. Ainsi, un centimorgan équivaut à 1% de recombinaison. Sur de courtes distances génétiques, cette unité de mesure a l'avantage d'être à peu près additive. Ceci n'est plus exact pour les marqueurs plus éloignés. En effet, lorsqu'un nombre pair d'enjambements se produit entre deux marqueurs, on n'observe pas de recombinaison. Puisque ceci se produit plus souvent pour des marqueurs éloignés, le taux de recombinaison a tendance à sous-estimer la distance génétique.

D'autres unités de mesures, physiques, peuvent être utilisées pour mesurer le génome. Par exemple, on peut compter le nombre de nucléotides, ou bases, séparant les marqueurs. Étant donné le grand nombre de bases constituant l'ADN, on utilise habituellement des multiples de mille. Ainsi, un kb (*kilo base*) est une unité comprenant mille

nucléotides. On peut établir un lien approximatif entre cette distance physique et le centimorgan. Chez l'homme, un centimorgan équivaut environ à un million de bases, c'est-à-dire 1 cM $\approx$ 1 000 kb.

1.3 Théorie de la coalescence

1.3.1 Principe de coalescence

Le principe de coalescence est un processus stochastique permettant de modéliser la généalogie de séquences d'ADN à rebours dans le temps. Nous en résumerons ici les grandes lignes de façon intuitive. Des preuves plus rigoureuses et des généralisations sont disponibles dans la littérature (Nordborg, 2001; Hein, Schierup et Wiuf, 2005). Dans sa forme la plus simple, le principe de coalescence s'applique à une population haploïde, sans recombinaison et de taille fixe. De plus, on suppose que les générations ne se côtoient pas, c'est-à-dire que tous les individus décèdent immédiatement après la naissance de la nouvelle génération. Enfin, chaque individu de l'échantillon choisit son ancêtre au hasard parmi ceux de la génération précédente selon un modèle d'urne avec remise. Ainsi, deux individus peuvent choisir le même ancêtre, ce qui donne lieu à la coalescence de deux lignées. Au bout de quelques générations, on en vient immanquablement à retrouver un ancêtre commun à tout l'échantillon, le MRCA (*Most Recent Commun Ancestor*). La figure 1.4 illustre ce que nous venons de décrire.

À chaque génération, la probabilité pour deux lignées en particulier de coalescer, c'est-à-dire de retrouver un ancêtre commun, est de $1/N$, où N est la taille de la population. La probabilité que ces deux lignées coalescent à la génération k est alors de $(1/N)(1 - 1/N)^{(k-1)}$. Ainsi, le temps avant coalescence de ces deux lignées suit une loi géométrique de moyenne N. Supposons que la taille de la population est très grande et considérons une échelle de temps où une unité correspond à N générations. On peut alors approximer le temps avant coalescence de deux lignées en particulier par une loi exponentielle de moyenne 1. En effet, si X suit une loi géométrique de moyenne N, $Y = X/N$ converge en loi vers une exponentielle de moyenne 1, lorsque N tend vers l'infini. De ce fait, le

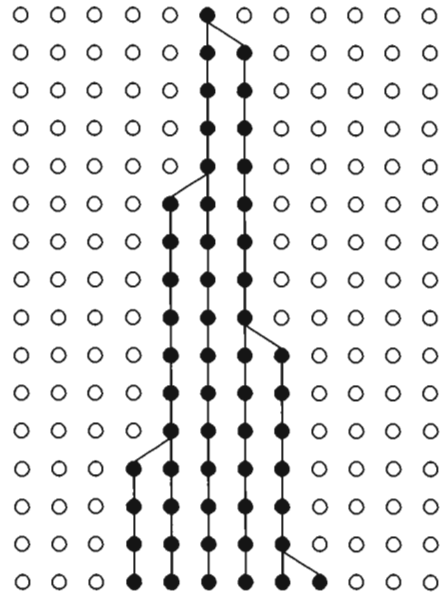

Figure 1.4 Généalogie d'un échantillon. Les points en noir représentent les 6 individus de l'échantillon (ligne inférieure) et leurs ancêtres sur 16 générations, pour une population de taille fixe N=12.

temps, qui était calculé en générations discrètes, devient une variable continue. Soit n le nombre de lignées présentes à un temps donné. Il y a au total $n(n-1)/2$ paires distinctes de lignées pouvant coalescer de manière indépendante. Le temps avant la prochaine coalescence entre deux de ces n lignées est le minimum de ces $n(n-1)/2$ temps de coalescence. Par conséquent, il peut être modélisé par une variable exponentielle de moyenne $2/[n(n-1)]$.

On en déduit un processus stochastique en temps continu qui modélise la généalogie des lignées de séquences haploïdes dans le passé par une suite d'événements de coalescence se produisant à des intervalles de temps décrits par une loi exponentielle. Ce processus peut être représenté par un arbre dont les noeuds sont les événements de coalescence et les branches correspondent aux intervalles de temps. Il est possible de considérer le phénomène des mutations neutres en les insérant simplement dans l'arbre selon un processus de poisson. Soit μ le taux de mutation individuel par génération. Considérons une

lignée en particulier, c'est-à-dire une branche de l'arbre. Les mutations se produiront sur celle-ci selon un processus de poisson de taux $\theta/2$, où $\theta = 2N\mu$. Le temps avant la prochaine mutation pour cette lignée est alors modélisé par une loi exponentielle de moyenne $2/\theta$. Pour chacune des n lignées présentes à un temps donné, les mutations surviennent indépendemment selon des processus de poisson. Ainsi, le temps avant la prochaine mutation est une variable exponentielle de moyenne $2/[n\theta]$. Lorsqu'une mutation survient, elle est ajoutée de manière aléatoire sur l'une des branches de l'arbre. La figure 1.5 représente un arbre de coalescence sur lequel on a inséré deux mutations.

1.3.2 Graphe de recombinaison ancestral

Le processus que nous venons de décrire s'applique à des séquences haploïdes sans recombinaison. Or, l'être humain est diploïde. De ce fait, les séquences génétiques qui forment son ADN se recombinent au moment de la reproduction. Le nombre de séquences d'ADN est donné par $2N$, où N est la taille de la population diploïde. Supposons dans un pre-

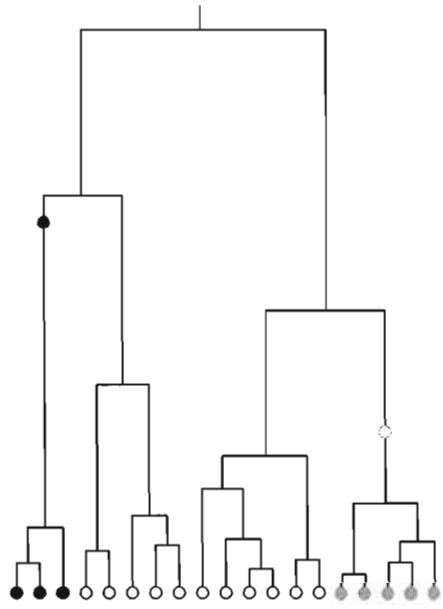

Figure 1.5 Arbre de coalescence. Les cercles noirs et gris représentent des mutations.

mier temps qu'il n'y a pas de recombinaison. Puisque nous modélisons la généalogie des séquences, et non des individus, la généralisation du principe à des individus diploïdes se résume dans ce cas à un changement d'échelle. Ainsi, nous retombons dans le cas précédent où chaque séquence choisit son parent de manière indépendante. L'échelle de temps sera toutefois notée en unités de $2N$ générations, soit la taille de la population de séquences haploïdes. Ainsi, ce simple changement d'échelle nous permet de décrire les temps avant coalescence, ou mutation, par les mêmes variables exponentielles que pour une population haploïde. Notons que le paramètre θ est calculé en fonction du nombre de séquences haploïdes, c'est-à-dire que $\theta = 4N\mu$.

Sans recombinaison, l'application du principe de coalescence à des individus diploïdes est donc directe. Or, les séquences d'ADN des humains recombinent. Pour cette raison, il faut généraliser le processus précédent afin d'y inclure cette réalité. Si on travaille à rebours dans le temps, une recombinaison consiste pour une séquence à se choisir deux séquences parents et à déterminer un point de recombinaison. Ainsi, tandis que chaque coalescence diminue le nombre de lignées, chaque recombinaison l'augmente de un. Soit r le taux individuel de recombinaison par génération. La probabilité qu'une lignée présente une première recombinaison à la génération k est alors de $r(1-r)^{(k-1)}$. Ainsi, le temps avant recombinaison suit une loi géométrique de moyenne $1/r$. En considérant une seule lignée et en ramenant l'échelle de temps à $2N$ générations, le temps avant recombinaison peut être approximé par une variable exponentielle de moyenne $2/\rho$, où $\rho = 4Nr$. Puisque l'on suppose que la population est très grande par rapport à l'échantillon, une recombinaison impliquant deux lignées de l'échantillon est un événement tellement rare qu'il peut être négligé. De ce fait, on peut supposer que les recombinaisons se produisent de manière indépendante sur les n lignées présentes à un temps donné. Le temps avant la prochaine recombinaison est alors modélisé par une variable exponentielle de moyenne $2/[n\rho]$.

Ce nouveau processus (Griffiths et Marjoram, 1996, 1997) est représenté par un graphe plutôt qu'un arbre. Il s'agit d'un processus de naissance et de mort pour lequel une naissance correspond à l'ajout d'une lignée à rebours dans le temps, c'est-à-dire une

recombinaison. Les morts correspondent quand à elles aux événements de coalescence. Les mutations sont finalement ajoutées sur les branches du graphe. Remarquons que le taux de naissance est linéaire en n, tandis que le taux de mort est quadratique, ce qui implique que le processus atteindra éventuellement un état où il ne restera plus qu'une seule séquence. Le matériel génétique hérité de cet ancêtre par ses descendants, incluant les formes dérivées obtenues par mutation, est nommé ancestral, par opposition au matériel génétique introduit par la recombinaison. Il est à noter qu'une recombinaison crée une insertion temporaire de matériel non ancestral dans la généalogie des individus, ce matériel génétique n'étant présent ni chez les individus de l'échantillon, ni chez le MRCA. En fait, chaque recombinaison crée une bifurcation dans une lignée, c'est-à-dire un point où deux segments du matériel génétique de celle-ci suivent des généalogies différentes.

La figure 1.6 illustre un graphe de recombinaison ancestral avec deux recombinaisons. Une application du graphe de recombinaison ancestral consiste à générer des généalogies compatibles avec un échantillon de séquences d'ADN observées aujourd'hui. Il permet aussi d'évaluer la vraisemblance de cet échantillon en conditionnant sur les généalogies possibles (Griffiths et Marjoram, 1996).

Le processus de coalescence que nous venons de décrire permet la modélisation des phénomènes de mutations et de recombinaisons à l'origine des variations dans le bagage génétique des populations. L'identification et l'étude de ces variations qui balisent le génome est l'objet de nombreuses recherches. Par exemple, la cartographie génétique vise à déterminer la position des gènes qui sont responsables de l'expression des caractères héréditaires. Nous avons présenté dans ce chapitre les concepts biologiques et statistiques essentiels à la compréhension des problématiques traitées dans ce travail. Nous verrons entre autres au prochain chapitre comment la modélisation de généalogies par le processus de coalescence peut être utilisée dans le contexte de la cartographie génétique.

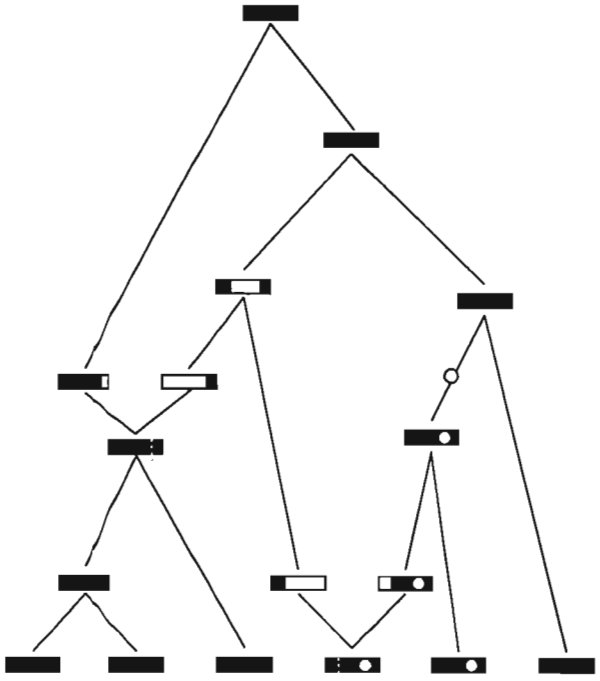

Figure 1.6 Graphe de coalescence ancestral. Les rectangles représentent les séquences génétiques. La partie noire est le matériel ancestral, tandis que la partie blanche représente le matériel non ancestral introduit par recombinaison. Le cercle représente une mutation.

CHAPITRE II

CARTOGRAPHIE GÉNÉTIQUE : LA MÉTHODE MAPARG

La science moderne a permis de mettre en lumière les principes de l'hérédité par l'étude de la génétique. Bien que beaucoup de travail ait été accompli, il demeure que la plupart des gènes humains n'ont pas encore été trouvés. L'identification des gènes impliqués dans l'expression des caractères héréditaires ne se résume pas à recueillir des profils génétiques. Il est aussi nécessaire d'analyser les données recueillies afin de prédire la position probable des gènes d'intérêt. Plusieurs méthodes statistiques de cartographie génétique ont été élaborée dans ce but. Nous décrirons dans ce chapitre une méthode de cartographie génétique basée sur le graphe de recombinaison ancestral et développée dans Larribe, Lessard et Schork (2002). Nous en présenterons la méthodologie ainsi que les distributions de probabilité impliquées. Enfin, nous verrons comment les auteurs de la méthode ont réussi à accélérer celle-ci par l'utilisation d'une vraisemblance composite. Notons qu'un résumé des notations utilisées dans ce chapitre est présenté à l'annexe B.

2.1 Méthodologie

2.1.1 Présentation

La méthode MapArg a été développée dans l'optique d'optimiser l'utilisation de l'information génétique fournie en faisant appel à un modèle mathématique reconnu ; le graphe de recombinaison ancestral. Pour ce faire, un modèle d'échantillonnage pondéré sur les généalogies de séquences est utilisé afin d'estimer la vraisemblance de la posi-

tion d'une mutation. Comme d'autres méthodes de cartographie (McPeek et Strahs, 1999; Morris, Whittaker et Balding, 2000; Zöllner et Pritchard, 2005), celle-ci suppose en premier lieu l'utilisation d'un échantillon de séquences haploïdes. Nous travaillerons dans les chapitres subséquents à généraliser la méthode de manière à intégrer la réalité diploïde et en permettre l'application à des populations humaines. Pour cette raison, la présentation théorique de la méthode diffère quelque peu de celle de l'article original, de manière à mettre en évidence le modèle d'échantillonnage pondéré.

Soit H_0, un échantillon d'haplotypes génotypés sur L marqueurs, dont $L-1$ ont une position connue et serviront de référence. Le marqueur dont la position est inconnue sera identifié comme étant la mutation causale d'un caractère dichotomique. Nous décrirons ici le cas de marqueurs de type SNP binaires, bien que la méthode puisse prendre en compte d'autres types de marqueurs. Nous considérerons une courte séquence génétique de longueur totale r, de sorte que les distances mesurées en Morgans soient additives. De plus, nous supposerons que la mutation causale est située entre le premier et le dernier marqueur de référence. Soit x_i la position du marqueur i, pour $i = 1, \ldots, L$. Afin de simplifier les calculs, posons $x_1 = 0$ et $x_1 < x_2 < \cdots < x_L$, ce qui implique que $x_L = r$. Enfin, notons par r_m la distance entre les marqueurs m et $m+1$, pour $m = 1, ..., L-1$. Ainsi, $x_m = \sum_{k=1,\ldots,m-1} r_k$, pour $2 \leq m \leq L$. Finalement, soit r_T, la position de la mutation causale. L'objectif de la méthode est d'estimer la vraisemblance $L(r_T) \equiv Q_{r_T}(H_0) \equiv Q(H_0 \mid r_T)$ de cette position.

La méthode consiste à estimer la vraisemblance en insérant successivement la mutation au centre de chacun des $L-2$ intervalles compris entre deux marqueurs de référence consécutifs. Ainsi, supposons que la mutation est insérée entre deux de ces marqueurs, de sorte qu'elle occupe maintenant le rang m où $2 \leq m \leq L-1$. La figure 2.1 (adaptée de Larribe, Lessard et Schork 2002) illustre la configuration de la séquence dans un pareil cas. La vraisemblance de la position est alors estimée sur cet intervalle par un modèle d'échantillonnage pondéré, la position centrale ($r_{m-1} = r_m$) étant utilisée comme valeur conductrice r_{T_0} pour la simulation.

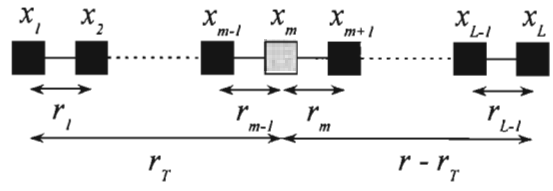

Figure 2.1 Configuration de la séquence.

2.1.2 Modèle d'échantillonnage pondéré

Considérons le graphe de recombinaison ancestral dont les états, à rebours dans le temps, sont donnés par H_0, H_1,..., $H_{\tau^\star-1}$, $H_{\tau^\star}$, où $H_{\tau^\star}$, supposé connu, est l'ancêtre commun de tous les haplotypes formant l'échantillon H_0 observé aujourd'hui. Chaque état H_i est donc formé de l'ensemble des séquences présentes à l'étape i, une nouvelle étape étant déterminée par un événement de coalescence, de recombinaison ou de mutation. En partant de l'ancêtre commun, on peut déduire la probabilité de cette généalogie, conditionnellement au paramètre r_T. En effet, la probabilité de chaque état ne dépend du passé que par le précédent. Ainsi, on obtient l'équation de récurrence

$$Q_{r_T}(H_\tau, H_{\tau+1}, ..., H_{\tau^\star}) \;=\; Q_{r_T}(H_\tau \mid H_{\tau+1}) Q_{r_T}(H_{\tau+1}, H_{\tau+2}, ..., H_{\tau^\star}),$$

ce qui nous donne

$$Q_{r_T}(H_0, H_1, ..., H_{\tau^\star}) \;=\; Q(H_{\tau^\star}) \prod_{\tau=0}^{\tau^\star-1} Q_{r_T}(H_\tau \mid H_{\tau+1}).$$

Puisque plusieurs généalogies peuvent donner lieu à l'échantillon observé H_0, on peut déduire la probabilité de celui-ci en considérant tous les graphes possibles. Notons que nous supposons ici que l'ancêtre commun est connu et unique, ce qui implique que $Q(H_{\tau^\star}) = 1$. On obtient la récurrence suivante en sommant sur tous les états possibles, une étape à rebours dans le temps,

$$Q_{r_T}(H_\tau) \;=\; \sum_{H_{\tau+1}} Q_{r_T}(H_\tau \mid H_{\tau+1}) Q_{r_T}(H_{\tau+1}). \tag{2.1}$$

Ce qui nous donne la probabilité de H_0,

$$Q_{r_T}(H_0) = \sum_{H_1}\sum_{H_2},\ldots,\sum_{H_{\tau^*-1}} Q(H_{\tau^*}) \prod_{\tau=0}^{\tau^*-1} Q_{r_T}(H_\tau \mid H_{\tau+1})$$

Si on pouvait énumérer tous les états possibles, on pourrait ainsi calculer la vraisemblance de r_T. En réalité, ce calcul est impraticable, puisque l'espace sur les graphes est infini.

Une façon de contourner ce problème consiste à générer un ensemble fini de graphes selon une distribution proposée $P_{r_{T_0}}$ correspondant à une valeur conductrice r_{T_0}. La vraisemblance recherchée sera alors estimée par une moyenne sur cette distribution. Ainsi, on peut réécrire l'équation de récurrence (2.1) sous la forme

$$\begin{aligned} Q_{r_T}(H_\tau) &= \sum_{H_{\tau+1}} \frac{Q_{r_T}(H_\tau \mid H_{\tau+1})}{P_{r_{T_0}}(H_{\tau+1} \mid H_\tau)} P_{r_{T_0}}(H_{\tau+1} \mid H_\tau) Q_{r_T}(H_{\tau+1}) \\ &= \sum_{H_{\tau+1}} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) P_{r_{T_0}}(H_{\tau+1} \mid H_\tau) Q_{r_T}(H_{\tau+1}), \end{aligned}$$

où

$$h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) = \frac{Q_{r_T}(H_\tau \mid H_{\tau+1})}{P_{r_{T_0}}(H_{\tau+1} \mid H_\tau)}.$$

Cette équation nous permet de réécrire la vraisemblance recherchée sous la forme d'une espérance sur la distribution $P_{r_{T_0}}$,

$$\begin{aligned} Q_{r_T}(H_0) &= \sum_{H_1}\sum_{H_2},\ldots,\sum_{H_{\tau^*-1}} Q(H_{\tau^*}) \prod_{\tau=0}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) P_{r_{T_0}}(H_{\tau+1} \mid H_\tau) \\ &= \sum_{H_1}\sum_{H_2},\ldots,\sum_{H_{\tau^*-1}} \left[Q(H_{\tau^*}) \prod_{\tau=0}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) \right] \left[\prod_{\tau=0}^{\tau^*-1} P_{r_{T_0}}(H_{\tau+1} \mid H_\tau) \right] \\ &= E_{P_{r_{T_0}}} \left[Q(H_{\tau^*}) \prod_{\tau=0}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) \right]. \end{aligned}$$

Rappelons que $Q(H_{\tau^*}) = 1$, l'ancêtre commun étant supposé connu et unique, ce qui nous permet de négliger ce facteur dans les calculs subséquents. Cette notation sous forme d'espérance nous permet d'estimer la vraisemblance $L(r_T)$ de la position de la mutation causale en calculant une moyenne sur les graphes générés. Ainsi, en simulant un

certain nombre de graphes selon la distribution $P_{r_{T_0}}$, on peut estimer la vraisemblance recherchée par une moyenne des valeurs de $\prod_{\tau=0}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1})$. Soit K le nombre de graphes, cette moyenne s'exprime alors sous la forme :

$$\begin{aligned} \hat{L}(r_T) &= \hat{Q}_{r_T}(H_0) \\ &= \frac{1}{K}\sum_{k=1}^{K}\left[\prod_{\tau=0}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau^k, H_{\tau+1}^k)\right]. \end{aligned}$$

Nous venons de décrire comment estimer la vraisemblance par une moyenne sur les graphes générés. Afin d'appliquer cette méthode, il est nécessaire de connaître les distributions Q et P.

2.2 Distributions

2.2.1 Événements

Les distributions que nous présentons ici sont celles présentées par Larribe et Lessard (2008) dans le cadre du principe de coalescence avec recombinaison et inspirées des travaux de Griffiths et Marjoram (1996).

Rappelons qu'à chaque étape du processus, un type d'événement parmi trois possibilités peut survenir à rebours dans le temps, une coalescence (C), une mutation (M) ou une recombinaison (R). Reprenons les notations suivantes sur les événements possibles :

1. C_i pour une coalescence de deux séquences de type i,
2. C_{ij}^k pour une coalescence de séquences de types i et j $(i \neq j)$ vers une séquence parentale de type k,
3. $M_i^j(m)$ pour une mutation au marqueur m d'une séquence de type i vers une séquence parentale de type j,
4. $R_i^{jk}(m)$ pour une recombinaison, dans l'intervalle m, d'une séquence de type i vers deux séquences parentales de types j et k.

Une coalescence peut se produire entre deux séquences dont le matériel ancestral est commun. Rappelons que le matériel ancestral est hérité directement de l'ancêtre com-

mun, par opposition au matériel introduit par la recombinaison. Ainsi, deux séquences identiques (i) peuvent toujours coalescer, ce qui correspond au premier cas. Lorsque la différence entre deux séquences (i et j) se limite à des insertions de matériel non ancestral, elles peuvent coalescer, ce qui correspond au deuxième cas. Le résultat de cette coalescence est une séquence parentale (k) réunissant tout le matériel ancestral des deux séquences.

Dans le modèle que nous étudions ici, on suppose que les mutations sont rares, c'est-à-dire que chacune ne survient qu'une seule fois dans la généalogie. Pour cette raison, une mutation ne peut survenir à un marqueur (m) que s'il existe une unique séquence (i) présentant une mutation à ce marqueur. Ceci correspond au troisième événement, et le résultat est une nouvelle séquence (j) identique à la première, sauf au marqueur considéré. Enfin, une recombinaison se produisant sur du matériel non ancestral ne sera pas considérée, puisqu'un tel événement n'apportera pas d'information pertinente sur la généalogie. Pour cette raison, une recombinaison peut se produire sur une séquence (i) à n'importe quel intervalle (m) situé entre deux marqueurs ancestraux. La figure 2.2 illustre chaque type d'événement par un exemple.

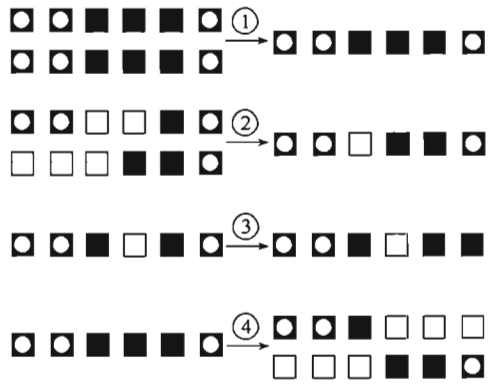

Figure 2.2 Exemples d'événements. On retrouve, dans l'ordre, une coalescence identique (1), une coalescence non identique (2), une mutation au sixième marqueur (3) et une recombinaison au troisième intervalle (4). Chaque carré est un marqueur. Un carré noir est ancestral, un cercle blanc représente une mutation et un carré blanc est non ancestral.

2.2.2 Probabilités et coalescence

Avant de décrire les probabilités des divers événements, il nous faut d'abord déterminer les taux de mutation et de recombinaison sur le matériel ancestral. Soit μ_m le taux de mutation par génération au marqueur m, pour $m = 1, ..., L$ et $\theta_m = 4N\mu_m$, le taux à l'échelle de coalescence. Le taux de mutation sur l'ensemble de la région chromosomique est alors donné par $\theta = \sum_m \theta_m$. De manière analogue, le taux de recombinaison entre les marqueurs m et $m+1$ à l'échelle de coalescence est donné par $\rho_m = 4Nr_m$, pour $m = 1, ..., L-1$. On en déduit le taux de recombinaison sur la globalité de la séquence, $\rho = \sum_m \rho_m$. Puisque les distances dépendent de la position de la mutation causale dans l'intervalle, nous noterons plutôt les taux de recombinaison comme fonctions de r_T. Ainsi, $\rho_m(r_T)$, pour $m = 1, ..., L-1$, représentent les taux de recombinaison obtenus en positionnant la mutation causale à r_T. Remarquons que le taux global de recombinaison ne dépend pas de r_T.

Soit n_i le nombre de séquences de type i à une étape τ du graphe, et $n = \sum_i n_i$. On notera par $m \in A^{(i)}$ un marqueur dont l'allèle est ancestral sur la séquence i. De même, on notera par $m \in B^{(i)}$ un intervalle de cette séquence compris entre deux marqueurs ancestraux, c'est-à-dire qu'il existe des valeurs $m_1 \leq m$ et $m_2 > m$ telles que $m_1 \in A^{(i)}$ et $m_2 \in A^{(i)}$. Cette notation nous permet de décrire le taux de mutation sur l'ensemble du matériel ancestral :

$$\alpha\theta n/2,$$

où

$$\alpha = \frac{2}{n\theta} \sum_i \left[\frac{n_i}{n} \sum_{m \in A^{(i)}} \theta_m \right].$$

On décrit aussi le taux de recombinaison sur le matériel ancestral :

$$\beta(r_T)\rho n/2,$$

où

$$\beta(r_T) = \frac{2}{n\rho} \sum_i \left[\frac{n_i}{n} \sum_{m \in B^{(i)}} \rho_m(r_T) \right].$$

Le taux de coalescence ne dépend quand à lui que de n et est donné par

$$n(n-1)/2.$$

Le prochain événement sera donc une coalescence, une mutation ou une recombinaison avec probabilités respectives,

$$\begin{aligned} P_\tau(C) &= \frac{n-1}{n-1+\alpha\theta+\beta(r_T)\rho}, \\ P_\tau(M) &= \frac{\alpha\theta}{n-1+\alpha\theta+\beta(r_T)\rho}, \\ P_\tau(R) &= \frac{\beta(r_T)\rho}{n-1+\alpha\theta+\beta(r_T)\rho}. \end{aligned}$$

Les probabilités précédentes, dérivées du processus de coalescence, peuvent être utilisées afin d'obtenir les probabilités dans le sens chronologique. D'abord, notons qu'il y a un nombre fini d'états $H_{\tau+1}$ ayant pu donner lieu à H_τ, chacun correspondant à un événement de coalescence, de mutation ou de recombinaison. Étant donné le type d'événement, il est possible de calculer la probabilité $Q_{r_T}(H_\tau \mid H_{\tau+1})$ en considérant les fréquences des séquences à l'étape $H_{\tau+1}$. Par exemple, si un événement de coalescence a eu lieu, il sera identique de type i avec probabilité $(n_i - 1)/(n-1)$. En effet, cela implique qu'à l'étape $H_{\tau+1}$, il y avait une séquence de type i en moins, soit $n_i - 1$, pour un total de $n-1$ séquences. La probabilité d'obtenir H_τ de $H_{\tau+1} = H_\tau + C_i$ est alors de $P\tau(C)(n_i-1)/(n-1)$, soit la probabilité que deux séquences de type i coalescent. Précisons que la notation $H_\tau + C_i$ représente un état $H_{\tau+1}$ dont la configuration est identique à H_τ, à une coalescence de séquences de type i près. On retrouve les autres probabilités de manière analogue, en regardant une étape à rebours dans le temps. Des détails supplémentaires sont donnés en annexe, page 145. La probabilité conditionnelle est alors donnée par :

$$Q_{r_T}(H_\tau \mid H_{\tau+1}) = \begin{cases} P_\tau(C)\frac{(n_i-1)}{(n-1)}, & \text{si } H_{\tau+1} = H_\tau + C_i, \\ P_\tau(C)\frac{(n_k+1-\delta_{ik}-\delta_{jk})}{(n-1)}, & \text{si } H_{\tau+1} = H_\tau + C_{ij}^k \ (i \neq j), \\ P_\tau(M)\frac{\theta_m}{\alpha\theta}\frac{(n_j+1)}{n}, & \text{si } H_{\tau+1} = H_\tau + M_i^j(m), \\ P_\tau(R)\frac{\rho_m(r_T)}{\beta(r_T)\rho}\frac{(n_j+1)(n_k+1)}{n(n+1)}, & \text{si } H_{\tau+1} = H_\tau + R_i^{jk}(m), \end{cases}$$

où $\delta_{ik} = 1$ si $i = k$ et 0 sinon. L'équation de récurrence (2.1) devient alors

$$\begin{aligned} Qr_T(H_\tau) &= P_\tau(C) \sum_{n_i>1} \frac{(n_i-1)}{(n-1)} Q(H_\tau + C_i) \\ &+ P_\tau(C) \sum_{\substack{i\neq j \\ compatible}} \frac{(n_k+1-\delta_{ik}-\delta_{jk})}{(n-1)} Q(H_\tau + C_{ij}^k) \\ &+ P_\tau(M) \sum_i \sum_{\substack{m\in A^{(i)} \\ unique}} \frac{\theta_m}{\alpha\theta} \frac{(n_j+1)}{n} Q(H_\tau + M_i^j(m)) \\ &+ P_\tau(R) \sum_i \sum_{m\in B^{(i)}} \frac{\rho_m(r_T)}{\beta(r_T)\rho} \frac{(n_j+1)(n_k+1)}{n(n+1)} Q(H_\tau + R_i^{jk}(m)). \end{aligned}$$

2.2.3 Distribution proposée

Le modèle d'échantillonnage pondéré nécessite une distribution permettant de générer des graphes, conditionnellement à une valeur conductrice r_{T_0}. Cette distribution peut être choisie de manière empirique. Idéalement, celle-ci sera le plus près possible de la vraie distribution. Cette dernière n'étant pas connue à rebours dans le temps, les auteurs ont choisi :

$$P_{r_{T_0}}(H_{\tau+1} \mid H_\tau) = \frac{b(H_\tau, H_{\tau+1})}{f(H_\tau)},$$

où

$$b(H_\tau, H_{\tau+1}) = \begin{cases} P_\tau(R) \frac{\rho_m(r_{T_0})}{\beta(r_{T_0})\rho n(n+1)} & \text{si } H_{\tau+1} = H_\tau + R_i^{jk}(m), \\ Q_{r_{T_0}}(H_\tau \mid H_{\tau+1}) & \text{sinon,} \end{cases}$$

et

$$f(H_\tau) = \sum_{H_{\tau+1}} b(H_\tau, H_{\tau+1})$$

est une constante de normalisation. En appliquant ces résultats au calcul de $h_{r_T r_{T_0}}(H_\tau, H_{\tau+1})$, on obtient :

$$\begin{aligned} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) &= \frac{Q_{r_T}(H_\tau \mid H_{\tau+1})}{P_{r_{T_0}}(H_{\tau+1} \mid H_\tau)} \\ &= \begin{cases} f(H_\tau) \frac{\rho_m(r_T)\beta(r_{T_0})}{\rho_m(r_{T_0})\beta(r_T)} (n_j+1)(n_k+1), & \text{si } H_{\tau+1} = H_\tau + R_i^{jk}(m); \\ f(H_\tau), & \text{sinon.} \end{cases} \end{aligned} \tag{2.2}$$

2.2.4 L'algorithme MapArg

Dans les sections précédentes, nous avons décrit les fondements et étapes de la méthode MapArg. Nous résumons dans cette section les grandes lignes de la méthode sous la forme d'un algorithme.

Algorithme MapArg

1. Choisir un ensemble fini de valeurs candidates r_T pour lesquelles la vraisemblance sera évaluée. Notons que la vraisemblance calculée n'est pas définie aux marqueurs de référence.
2. Pour chaque intervalle m compris entre deux marqueurs de référence :
   (a) Construire la liste H_0 des séquences observées en insérant le marqueur de la mutation causale au centre de l'intervalle, de sorte qu'il se retrouve au rang m, tel qu'illustré à la figure 2.1 de la page 19. La position centrale r_{T_0} servira de valeur conductrice.
   (b) Pour chacun des K graphes à construire :
   - Pour chaque étape τ telle que $0 \leq \tau \leq \tau^* - 1$:
      i. Déterminer les états $H_{\tau+1}$ admissibles en considérant tous les événements possibles, tels que décrits à la section 2.2.1, page 21.
      ii. Pour chaque état, déterminer la probabilité associée en évaluant $b(H_\tau, H_{\tau+1})/f(H_\tau)$ (voir section 2.2.3, page 25).
      iii. Générer le nouvel état $H_{\tau+1}$ selon la distribution calculée à l'étape précédente.
      iv. Pour chaque valeur de r_T comprise dans l'intervalle, évaluer la fonction $h_{r_T r_{T_0}}(H_\tau, H_{\tau+1})$ décrite à l'équation (2.2), page 25.
   -Pour chaque valeur de r_T, calculer le produit $\prod_{\tau=0}^{\tau^*-1} h_{r_T, r_{T_0}}(H_\tau, H_{\tau+1})$.
   (c) Pour chaque position r_T de l'intervalle, calculer la moyenne sur tous les graphes :

$$\hat{Q}_{r_T}(H_0) = \frac{1}{K} \sum_{k=1}^{K} \left[\prod_{\tau=0}^{\tau^*-1} h_{r_T, r_{T_0}}(H_\tau^k, H_{\tau+1}^k) \right].$$

2.3 Vraisemblance composite conditionnelle

2.3.1 Vraisemblance composite

L'algorithme précédent est basé sur un modèle d'échantillonnage pondéré nécessitant la simulation de graphes de recombinaison ancestraux. De ce fait, le temps d'estimation augmente de manière importante avec le nombre de marqueurs et la longueur de la séquence. En effet, plus le nombre de marqueurs est grand, plus le nombre d'événements dans la généalogie risque d'être important et plus la simulation est ardue. Aussi, plus la longueur est importante, plus les événements de recombinaison sont probables, ce qui augmente le temps avant d'atteindre l'ancêtre commun. Il peut alors s'avérer utile de réduire la simulation à de plus petites séquences.

Une vraisemblance composite est une «pseudo-vraisemblance» formée de la combinaison des vraisemblances marginales obtenues sur des sous-ensembles des données originales. Il s'agit d'une approximation de la vraisemblance cherchée, qui dépend des sous-ensembles choisis. Reprenons la définition de Varin (2008). Soit Y, une variable aléatoire multidimensionnelle sur un espace $\mathcal{Y}$, et $\theta \in \Theta$, un paramètre inconnu. Supposons que l'on souhaite évaluer la vraisemblance $L(\theta) = f(y \mid \theta)$, basée sur une observation $y \in \mathcal{Y}$, mais que celle-ci est difficile à évaluer. Considérons dans ce cas un ensemble d'événements $\{A_j, j = 1, ..., J\}$, de telle sorte que les distributions $f(y, A_j \mid \theta)$ soient plus aisées à obtenir. La vraisemblance composite est alors définie comme le produit pondéré de ces distributions :

$$CL(\theta) = \prod_j f(y, A_j \mid \theta)^{\omega_i}, \tag{2.3}$$

où ω_i est un poids positif. Le résultat obtenu, généralement plus facile à calculer, est une approximation de la vraisemblance complète. Elle est de plus en plus utilisée dans les contextes où les bases de données sont très vastes et les calculs exacts ardus.

Il existe plusieurs façons de former des sous-ensembles à partir des données initiales. Dans le cas qui nous intéresse, la réduction du nombre de marqueurs peut s'avérer utile. Pour cette raison, nous considérerons un découpage de la séquence d'origine en J

fenêtres de marqueurs, formant autant de sous-ensembles. Ainsi, on obtient un ensemble d'événements $\{A_j, 1 \leq j \leq J\}$, où A_j représente les marqueurs et intervalles de la fenêtre j. L'application directe de la définition (2.3), avec des poids unitaires, nous permet d'obtenir la vraisemblance composite d'une position r_T par le produit des distributions marginales :

$$CL(r_T) = \prod_j Q(H_0, A_j \mid r_T).$$

Notons que (H_0, A_j) représente ici l'information sur les haplotypes obtenue en retirant de H_0 tous les marqueurs de référence qui ne sont pas compris dans la fenêtre j.

2.3.2 Vraisemblance conditionnelle

Considérons dans un premier temps un découpage en fenêtres juxtaposées, de telle sorte que chaque intervalle entre marqueurs de référence est entièrement compris dans une unique fenêtre, tel qu'illustré à la figure 2.3. Supposons la mutation causale dans l'intervalle m, situé entre les marqueurs de référence $m-1$ et m. Notons par j_m l'unique fenêtre comprenant cet intervalle. Remarquons que la densité sur les autres fenêtres ne dépend pas de la mutation causale, c'est-à-dire que

$$Q(H_0, A_j \mid r_T) = Q(H_0^*, A_j), \text{ lorsque } j \neq j_m, \tag{2.4}$$

où H_0^* représente l'information sur les marqueurs de référence de H_0 sans la mutation causale.

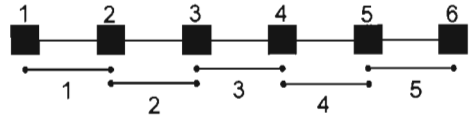

Figure 2.3 Découpage en fenêtres juxtaposées. Ici, on a 6 marqueurs de référence et 5 fenêtres de 2 marqueurs.

Il est alors possible d'exprimer la vraisemblance composite donnée à l'équation (2.4) comme une fonction définie par morceaux sur les $L-2$ intervalles délimités par les $L-1$ marqueurs de référence et pouvant contenir la mutation causale. Pour chaque intervalle m, on peut exprimer la vraisemblance composite de la position r_T définie sur celui-ci par

$$CL_m(r_T) = Q(H_0, A_{j_m} \mid r_T) \prod_{j \neq j_m} Q(H_0^*, A_j).$$

Remarquons que le produit $\prod_j Q(H_0^*, A_j)$ est un facteur qui ne dépend que du découpage des fenêtres, et non de la position de la mutation causale. Il s'agit donc d'une constante multiplicative qui n'intervient pas dans l'estimation de r_T. On peut mettre celui-ci en évidence en écrivant l'expression précédente sous la forme

$$CL_m(r_T) = \frac{Q(H_0, A_{j_m} \mid r_T)}{Q(H_0^*, A_{j_m})} \prod_j Q(H_0^*, A_j).$$

Il ne reste finalement qu'à évaluer le quotient $Q(H_0, A_{j_m} \mid r_T)/Q(H_0^*, A_{j_m})$, qui peut être vu comme une vraisemblance conditionnelle à la fenêtre. Pour chaque intervalle m et chaque fenêtre j, définissons la fonction

$$L_{m,j}(r_T \mid H_0^*, A_j) = \begin{cases} \frac{Q(H_0, A_j \mid r_T)}{Q(H_0^*, A_j)}, & \text{si } r_T \text{ est inclus dans l'intervalle } m \text{ de la fenêtre } j \\ 1, & \text{sinon.} \end{cases}$$

On obtient la vraisemblance composite conditionnelle sur l'ensemble de la séquence en effectuant le produit des vraisemblances conditionnelles sur chaque intervalle :

$$CCL(r_T) = \prod_{m=2}^{L-1} L_{m,j_m}(r_T \mid H_0^*, A_{j_m}).$$

Généralisons ce que nous venons de décrire. Considérons maintenant un découpage de la séquence d'origine en $J = L - d$ fenêtres de d marqueurs, superposées de manière à ce qu'il y ait une différence de un marqueur entre deux fenêtres consécutives, tel qu'illustré à la figure 2.4. Dans un tel découpage, l'intervalle m est inclus dans toutes les fenêtres j telles que $\underline{j}(m) \leq j \leq \overline{j}(m)$, où

$$\begin{aligned} \underline{j}(m) &= \max(1, m+1-d), \\ \overline{j}(m) &= \min(m-1, L-d). \end{aligned}$$

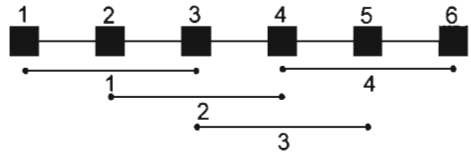

Figure 2.4 Découpage en fenêtres superposées. On a ici 6 marqueurs de référence et 4 fenêtres de 3 marqueurs.

Cet intervalle est donc inclus dans $\bar{j}(m) - \underline{j}(m) + 1$ fenêtres. Ainsi, le simple produit des vraisemblances accordera un poids différent à chaque intervalle, causant un biais dans l'estimation. Pour corriger ce problème, il suffit d'effectuer une moyenne géométrique des vraisemblances conditionnelles. On obtient alors

$$CCL_d(r_T) = \prod_{m=2}^{L-1} \left(\prod_{j=\underline{j}(m)}^{\bar{j}(m)} L_{m,j}(r_T \mid H_0^*, A_j) \right)^{\omega_m}, \tag{2.5}$$

où

$$\omega_m = \frac{1}{\bar{j}(m) - \underline{j}(m) + 1}.$$

Cette vraisemblance composite conditionnelle est une approximation de la vraisemblance sur les données complètes, à un facteur $\prod_j Q(H_0^*, A_j)$ près qui dépend du découpage choisi. Notons que le calcul de $Q(H_0^*)$ se fait de manière similaire à celui de $Q(H_0)$. Il suffit de ne pas tenir compte de la mutation causale dans les calculs de probabilités.

2.3.3 Algorithme MapArg avec vraisemblance composite

Nous avons expliqué précédemment que l'algorithme MapArg original est difficile à appliquer lorsque la taille des séquences est très importante. Pour cette raison, nous avons décrit dans les pages précédentes une nouvelle version basée sur la vraisemblance composite. L'objectif de cette section est de résumer ce nouvel algorithme. Pour ce faire, nous présenterons d'abord une légère variante de l'algorithme original permettant de calculer la probabilité des marqueurs de référence, sans égard à la mutation causale.

Algorithme MapArg sur les marqueurs de référence

1. Construire H_0^* en ne considérant que les marqueurs de référence.
2. Pour chacun des K graphes à construire :
   - Pour chaque étape τ telle que $0 \leq \tau \leq \tau^* - 1$:
     (a) Déterminer les états $H_{\tau+1}$ admissibles, en considérant tous les événements possibles, tels que décrits à la section 2.2.1, page 21.
     (b) Pour chaque état, déterminer la probabilité associée en évaluant $b(H_\tau, H_{\tau+1})/f(H_\tau)$ (voir section 2.2.3, page 25). Notons qu'ici, les paramètres ne dépendent pas de r_{T_0}, puisque la mutation causale n'est pas considérée.
     (c) Générer le nouvel état $H_{\tau+1}$ selon la distribution calculée à l'étape précédente.
     (d) Évaluer la fonction

$$h(H_\tau, H_{\tau+1}) = \begin{cases} f(H_\tau)(n_j+1)(n_k+1), & \text{si } H_{\tau+1} = H_\tau + R_{ij}^k\,; \\ f(H_\tau), & \text{sinon.} \end{cases}$$

   - Calculer le produit $\prod_{\tau=0}^{\tau^*-1} h(H_\tau, H_{\tau+1})$.
3. Estimer la probabilité des données en calculant la moyenne sur tous les graphes :

$$\hat{Q}(H_0^*) = \frac{1}{K}\sum_{k=1}^{K}\left[\prod_{\tau=0}^{\tau^*-1} h(H_\tau, H_{\tau+1})\right].$$

Algorithme MapArg avec vraisemblance composite

1. Choisir un ensemble fini de valeurs candidates r_T pour lesquelles la vraisemblance sera évaluée. Notons que la vraisemblance calculée n'est pas définie aux marqueurs de référence.

2. Découper la séquence en J fenêtres de d marqueurs, tel qu'illustré à la figure 2.4 de la page 30.

3. Pour chaque intervalle m, énumérer les fenêtres qui le contiennent, en évaluant

$$\begin{aligned} \underline{j}(m) &= \max(1, m+1-d), \\ \overline{j}(m) &= \min(m-1, L-d). \end{aligned}$$

4. Calculer les poids

$$\omega_m = \frac{1}{\overline{j}(m) - \underline{j}(m) + 1},$$

pour $m = 2, ..., L-1$.

5. Pour chaque fenêtre j :

   (a) Créer A_j en ne conservant que les marqueurs et intervalles de la fenêtre j.

   (b) Évaluer $\hat{Q}(H_0^*, A_j)$ en appliquant l'algorithme MapArg sur les marqueurs de référence.

   (c) Pour chaque intervalle m tel que $\underline{j}(m) \leq j \leq \overline{j}(m)$:

      i. Évaluer $\hat{Q}(H_0, A_j)$ en appliquant l'étape 2 de l'algorithme MapArg original sur les marqueurs et intervalles de la fenêtre.

      ii. Pour chaque valeur de r_T, calculer la vraisemblance conditionnelle :

$$\hat{L}_{m,j}(r_T \mid H_0^*, A_j) = \begin{cases} \hat{Q}(H_0, A_j)/\hat{Q}(H_0^*, A_j), & \text{si } r_T \text{ est inclus dans } m\,; \\ 1, & \text{sinon.} \end{cases}$$

6. Pour chaque position r_T, combiner les vraisemblances conditionnelles en évaluant

$$\widehat{CCL}(r_T) = \prod_{m=2}^{L-1} \left(\prod_{j=\underline{j}(m)}^{\overline{j}(m)} \hat{L}_{m,j}(r_T \mid H_0^*, A_{j_m}) \right)^{\omega_m}.$$

Les algorithmes que nous avons présentés permettent d'appliquer la méthode MapArg à un échantillon de séquences haploïdes, dans le but d'estimer la position d'une mutation causale. On suppose pour ce faire la connaissance des haplotypes de chaque individu, de même que le statut au gène causal. Toutefois, les échantillons disponibles se présentent rarement sous cette forme. Cette problématique est traitée dans les chapitres suivants.

CHAPITRE III

ESTIMATION DES HAPLOTYPES

Nous avons présenté au chapitre précédent une méthode statistique de cartographie génétique pour laquelle on suppose la connaissance des haplotypes formant le génotype des individus échantillonnés. Cette hypothèse correspond toutefois rarement à la réalité. De manière générale, l'information génétique disponible est présentée sous la forme de génotypes non phasés, c'est-à-dire pour lesquels on ignore l'identité des deux haplotypes formant le diplotype. Afin de solutionner ce problème, plusieurs procédures statistiques ont été développées. Nous présenterons dans ce chapitre des algorithmes reconnus pour leurs performances ou pour leur contribution historique. Nous comparerons ensuite les méthodes présentées et les solutions proposées par certains auteurs de méthodes de cartographie génétique.

3.1 Principe de parcimonie

3.1.1 Présentation

Le principe de parcimonie de Clark (1990) est le premier algorithme statistique connu d'estimation des haplotypes. Comme son nom l'indique, il se base sur l'hypothèse que la variabilité du génome est relativement faible et, par conséquent, que le nombre d'haplotypes distincts doit être le plus petit possible.

L'algorithme doit avoir pour point de départ une liste d'haplotypes connus. Cette liste peut être obtenue en considérant les individus homozygotes. En théorie, l'algorithme

peut débuter avec un unique haplotype. En pratique, la connaissance d'un plus grand nombre peut être nécessaire. Pour chaque élément de la liste, on cherche alors à trouver les individus pour lesquels le génotype non résolu est compatible. Un haplotype h_i est compatible avec un génotype g s'il existe une séquence complémentaire h_j telle que le couple (h_i, h_j) forme un diplotype donnant lieu à l'observation de g. Lorsqu'un haplotype compatible est trouvé, on considère que le diplotype correspondant est la solution du génotype. Notons qu'on considérera en premier lieu des diplotypes formés de deux haplotypes de la liste. Autrement, la séquence complémentaire est alors ajoutée à la liste. On recommence l'opération jusqu'à ce que tous les génotypes soient résolus ou jusqu'à ce qu'il ne soit plus possible de continuer sans information supplémentaire.

L'avantage de cette méthode est qu'elle est simple à comprendre et à mettre en pratique. Cependant, les résultats obtenus dépendent grandement des informations de départ et de l'ordre selon lequel on parcourt la liste. De plus, il est aussi possible que l'algorithme ne permette pas de reconstruire les haplotypes pour un ou plusieurs individus. L'application du principe de parcimonie de Clark est limitée à un nombre restreint de marqueurs, sur de courtes distances génétiques. En effet, l'hypothèse d'un nombre limité d'haplotypes ne tient plus lorsque les séquences sont très longues et que la recombinaison devient plus importante.

L'algorithme de Clark n'est pas nécessairement la solution optimale au principe de parcimonie. En effet, bien que le principe semble simple, il n'est pas trivial d'en obtenir un algorithme qui utilise l'information de manière optimale. De nouvelles méthodes d'estimation basées sur ce principe ont été développées. Des résultats intéressants ont été obtenus sur des séquences de dix marqueurs (Huang, Chao et Ting, 2005).

3.1.2 Illustration

Appuyons la description précédente du principe de parcimonie par un exemple. Considérons l'échantillon de génotypes donné au tableau 3.1. Notons qu'ici la fréquence de chaque génotype n'a aucune incidence sur l'algorithme de Clark. En effet, seule la liste

Tableau 3.1 Liste de génotypes

Génotype *	Fréquence
$\{(a,a)(b,b)(c,c)\}$	3
$\{(a,A)(b,b)(c,C)\}$	4
$\{(a,A)(b,B)(c,c)\}$	1
$\{(A,A)(b,b)(C,C)\}$	2
$\{(A,A)(b,B)(c,C)\}$	1
total	11

*. Les trois marqueurs sont délimités par des parenthèses. Les allèles sont séparés par des virgules. Notons que l'ordre des allèles à un marqueur n'a pas d'importance ici, de sorte que (c,C) est équivalent à (C,c).

obtenue compte. En considérant les individus homozygotes, on obtient la liste d'haplotypes connus $\{abc, AbC\}$. Il reste alors trois génotypes à résoudre : $\{(a,A)(b,b)(c,C)\}$, $\{(a,A)(b,B)(c,c)\}$ et $\{(A,A)(b,B)(c,C)\}$. L'haplotype abc, conjointement avec AbC, est compatible avec le premier de ces génotypes. De même, il est aussi compatible avec le second, mais nécessite une séquence complémentaire ABc, qui est ajoutée à la liste. L'haplotype AbC, combiné à ABc, peut former le dernier génotype. On obtient finalement les reconstructions données au tableau 3.2. Notons ici que l'ordre des haplotypes formant un diplotype n'a pas d'importance.

Tableau 3.2 Résolution des génotypes

Génotype	Diplotype
$\{(a,a)(b,b)(c,c)\}$	(abc, abc)
$\{(\mathbf{a},\mathbf{A})(\mathbf{b},\mathbf{b})(\mathbf{c},\mathbf{C})\}$ *	$(\mathbf{abc}, \mathbf{AbC})$
$\{(\mathbf{a},\mathbf{A})(\mathbf{b},\mathbf{B})(\mathbf{c},\mathbf{c})\}$	$(\mathbf{abc}, \mathbf{ABc})$
$\{(A,A)(b,b)(C,C)\}$	(AbC, AbC)
$\{(\mathbf{A},\mathbf{A})(\mathbf{b},\mathbf{B})(\mathbf{c},\mathbf{C})\}$	$(\mathbf{AbC}, \mathbf{ABc})$

*. Les lignes en caractères gras identifient les génotypes ayant nécessité une résolution.

3.2 Méthode bayésienne

3.2.1 Échantillonnage de Gibbs

Il existe plusieurs méthodes bayésiennes d'estimation des haplotypes. En supposant une distribution *a priori* pour les diplotypes, l'objectif de ces méthodes est de décrire la distribution *a posteriori* $P(D \mid G)$, où G représente un échantillon de génotypes et D, les diplotypes correspondants. Parmi les méthodes bayésiennes, celle de Phase (Stephens et Donnelly, 2001), basée sur un échantillonnage de Gibbs, semble être actuellement la plus performante (Xu *et al.*, 2004 ; Stephens et Donnelly, 2003).

Étant donné la nature complexe des phénomènes génétiques et le nombre de paramètres impliqués, la distribution *a posteriori* est ardue, voire impossible à calculer. Une façon de contourner ce problème consiste à faire un échantillonnage de Gibbs, mieux connu sous le terme anglais de *Gibbs sampling*. Comme son nom l'indique, cet outil statistique permet d'obtenir un échantillon d'une distribution $P(X)$, où X est un vecteur aléatoire de dimension finie. Pour ce faire, il suffit que la distribution conditionnelle $P(X_i = x_i \mid X_{-i})$ d'un élément étant donné les autres soit connue. On génère alors une chaîne de Markov dont les états $X^{(k)}$ correspondent à des valeurs possibles du vecteur.

À chaque étape k, on choisit au hasard un élément X_i à estimer. On considère que les autres éléments sont connus et sont donnés par $X_{-i}^{(k-1)}$. Notons que $X^{(0)}$ est un point de départ aléatoire qui n'affectera pas le résultat final, mais uniquement la vitesse de convergence. On actualise ensuite X en attribuant à $X_i^{(k)}$ une valeur x_i générée selon la distribution conditionnelle $P(X_i = x_i|X_{-i}^{(k-1)})$. L'état $X^{(k)}$ s'obtient finalement en posant $X_{-i}^{(k)} = X_{-i}^{(k-1)}$, c'est-à-dire que les autres éléments sont conservés tels quels. La distribution stationnaire de la chaîne correspond alors à $P(X)$. On génère ainsi b itérations, où b est un entier jugé assez grand pour atteindre la distribution stationnaire. Une fois la convergence obtenue, chaque étape de la chaîne peut être considérée comme une observation tirée de $P(X)$. Pour générer des observations indépendantes, il suffit de retenir des étapes éloignées, par exemple en retenant un état à toutes les m itérations.

L'application de l'échantillonnage de Gibbs au problème des diplotypes se résume essentiellement à décrire la distribution conditionnelle $P(D_i = d_i \mid D_{-i}, G)$, où D_i est la paire d'haplotypes associée à l'individu i que l'on doit actualiser et D_{-i} représente les diplotypes précédemment attribués au reste de l'échantillon. Les hypothèses utilisées pour décrire cette distribution varient selon les auteurs, donnant lieu à de nouveaux algorithmes.

3.2.2 L'algorithme de Stephens et Donnelly : Phase

Nous avons mentionné que la construction d'un algorithme basé sur un échantillonnage de Gibbs suppose la connaissance de $P(D_i = d_i \mid D_{-i}, G)$, ce qui n'est pas le cas en général et représente un calcul impraticable pour la plupart des modèles utilisés en génétique. Toutefois, si on note $d_i = (h_{i,1}, h_{i,2})$, il est possible d'écrire

$$P(D_i = d_i \mid D_{-i}, G) \propto \pi(h_{i,1} \mid H)\pi(h_{i,2} \mid H),$$

où $\pi(h \mid H)$ est la probabilité d'obtenir un nouvel haplotype h étant donné que l'on a observé l'ensemble H des séquences formant les diplotypes D_{-i}. Notons qu'ici le diplotype d_i considéré doit être compatible avec le génotype de l'individu. Dans le cas contraire, on a directement que $P(D_i = d_i \mid D_{-i}, G) = 0$.

Dans un précédent article (Stephens et Donnelly, 2000), les auteurs de la méthode ont choisi de s'inspirer du principe de coalescence afin de décrire une approximation cohérente de $\pi(h \mid H)$. Soit E l'ensemble des haplotypes théoriquement possibles, r_α la fréquence des séquences de type α dans H et r le nombre d'haplotypes dans H. Notons par P la matrice des mutations, où $P_{\alpha h}$ représente la probabilité que la mutation d'une séquence de type α résulte en une séquence de type h. Ainsi, $P_{\alpha h}^{(s)}$ est la probabilité d'obtenir h de α après un nombre s de mutations. Mentionnons que P est supposée inversible. Selon le principe de coalescence, la prochaine séquence sera obtenue par mutation avec probabilité $\theta/(r+\theta)$. On modélise alors le phénomène des mutations en choisissant une séquence α au hasard parmi celles contenues dans H et en la faisant muter s fois selon une loi géométrique. En sommant sur toutes les séquences α disponibles, on obtient

$$\pi(h \mid H) = \sum_{\alpha \in E} \frac{r_\alpha}{r} \sum_{s=0}^{\infty} \left(\frac{\theta}{r+\theta}\right)^s \left(\frac{r}{r+\theta}\right) P_{\alpha h}^{(s)}.$$

L'expression précédente ne peut être évaluée directement. En effet, la dimension de la matrice P, ainsi que la somme infinie, rendent le calcul impraticable. Notons $\lambda = \theta/(r+\theta)$. On a alors

$$\begin{aligned} \pi(h \mid H) &= \sum_{\alpha \in E} \frac{r_\alpha}{r} \sum_{s=0}^{\infty} \lambda^s (1-\lambda) P_{\alpha h}^{(s)} \\ &= \sum_{\alpha \in E} \frac{r_\alpha}{r} (1-\lambda) \sum_{s=0}^{\infty} (\lambda P)_{\alpha h}^{(s)}. \end{aligned}$$

Puisque P est stochastique, toutes les valeurs propres de λP sont, en valeur absolue, inférieures à $\lambda < 1$. On a ainsi $\lim_{s \to \infty} (\lambda P)^{(s)} = \mathbf{0}$. La série géométrique converge et on se retrouve alors avec le résultat

$$\pi(h \mid D) = \sum_{\alpha \in E} \frac{r_\alpha}{r} M_{\alpha h}, \tag{3.1}$$

où $M = (1-\lambda)(I - \lambda P)^{-1}$ ne dépend pas de l'identité des séquences formant H.

Pour appliquer l'algorithme, il faut donc en premier lieu déterminer les paramètres du modèle de coalescence, c'est-à-dire θ et P, qui dépendent tous deux du contexte et des hypothèses retenues. Notons que le nombre total d'haplotypes r de H est constant à

chaque itération, même si l'identité des séquences varie. Ainsi, si n est le nombre total d'individus, D_{-1} en contient $n-1$, ce qui implique que $r = 2(n-1)$. La matrice M n'a donc besoin d'être calculée qu'une seule fois en utilisant ces paramètres. À chaque étape k, on choisit aléatoirement un individu i pour lequel le diplotype devra être estimé. On obtient $H^{(k-1)}$, ainsi que les fréquences r_α, en supposant les autres diplotypes connus. Ceci nous permet de calculer la probabilité associée à chaque diplotype $d_i = (h_{i,1}, h_{i,2})$, compatible avec le génotype de l'individu i, en combinant l'équation (3.1) et le fait que $P(D_i = d_i \mid D_{-i}^{(k-1)}, G) \propto \pi(h_{i,1} \mid H^{(k-1)})\pi(h_{i,2} \mid H^{(k-1)})$. On actualise enfin D en attribuant à $D_i^{(k)}$ un diplotype d_i généré selon la distribution $P(D_i = d_i \mid D_{-i}^{(k-1)}, G)$. Mentionnons qu'on ne peut prouver que la distribution conditionnelle que nous venons de décrire correspond réellement à une distribution conjointe $P(D \mid G)$. Pour cette raison, on dit que l'algorithme est en fait un «pseudo» échantillonnage de Gibbs.

Nous avons présenté ici la première version de Phase. Notons que d'autres versions ont été élaborées, l'une d'elles prenant en considération le phénomène des recombinaisons. L'algorithme a aussi été amélioré afin d'augmenter la vitesse de convergence. Le logiciel Phase 2.0 présentement distribué intègre ces modifications (Stephens et Scheet, 2005; Stephens et Donnelly, 2003).

3.2.3 Illustration

Reprenons l'exemple donné au tableau 3.1. Puisqu'il y a trois marqueurs présentant chacun deux allèles, il existe théoriquement huit haplotypes. Puisqu'il y a onze individus et que l'on en reconstruit un à chaque étape, il en reste dix dont les diplotypes sont supposés connus. On a alors directement que le nombre d'haplotypes connus est $r = 2 \times 10 = 20$. Le paramètre θ dépend des caractéristiques de la population. Pour les besoins de l'exemple, nous prendrons ici arbitrairement $\theta = 2$, ce qui nous donne $\lambda = 1/11$. Considérons la matrice de mutation P suivante, qui correspond à un modèle où chaque mutation implique un unique marqueur choisi au hasard. Afin d'appliquer l'algorithme, on calcule aussi la matrice $M = (1-\lambda)(I - \lambda P)^{-1}$.

$$P = \begin{array}{c} \\ \text{abc} \\ \text{abC} \\ \text{aBc} \\ \text{aBC} \\ \text{Abc} \\ \text{AbC} \\ \text{ABc} \\ \text{ABC} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{abc} & \text{abC} & \text{aBc} & \text{aBC} & \text{Abc} & \text{AbC} & \text{ABc} & \text{ABC} \end{array} \\ \begin{bmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 1/3 \\ 1/3 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 0 & 0 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \end{bmatrix} \end{array}$$

$$M = \begin{array}{c} \\ \text{abc} \\ \text{abC} \\ \text{aBc} \\ \text{aBC} \\ \text{Abc} \\ \text{AbC} \\ \text{ABc} \\ \text{ABC} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{abc} & \text{abC} & \text{aBc} & \text{aBC} & \text{Abc} & \text{AbC} & \text{ABc} & \text{ABC} \end{array} \\ \begin{bmatrix} 0{,}9116 & 0{,}0277 & 0{,}0277 & 0{,}0017 & 0{,}0277 & 0{,}0017 & 0{,}0017 & 0{,}0002 \\ 0{,}0277 & 0{,}9116 & 0{,}0017 & 0{,}0277 & 0{,}0017 & 0{,}0277 & 0{,}0002 & 0{,}0017 \\ 0{,}0277 & 0{,}0017 & 0{,}9116 & 0{,}0277 & 0{,}0017 & 0{,}0002 & 0{,}0277 & 0{,}0017 \\ 0{,}0017 & 0{,}0277 & 0{,}0277 & 0{,}9116 & 0{,}0002 & 0{,}0017 & 0{,}0017 & 0{,}0277 \\ 0{,}0277 & 0{,}0017 & 0{,}0017 & 0{,}0002 & 0{,}9116 & 0{,}0277 & 0{,}0277 & 0{,}0017 \\ 0{,}0017 & 0{,}0277 & 0{,}0002 & 0{,}0017 & 0{,}0277 & 0{,}9116 & 0{,}0017 & 0{,}0277 \\ 0{,}0017 & 0{,}0002 & 0{,}0277 & 0{,}0017 & 0{,}0277 & 0{,}0017 & 0{,}9116 & 0{,}0277 \\ 0{,}0002 & 0{,}0017 & 0{,}0017 & 0{,}0277 & 0{,}0017 & 0{,}0277 & 0{,}0277 & 0{,}0116 \end{bmatrix} \end{array}$$

Prenons comme point de départ une reconstruction aléatoire. Tout autre point de départ aurait pu être utilisé. Le tableau 3.3 donne les reconstructions initiales pour chaque individu. La première étape consiste à choisir un individu au hasard parmi ceux qui sont ambigus. Supposons que l'individu à mettre à jour est le cinquième (en caractères gras dans le tableau). À partir des reconstructions initiales pour les génotypes restants, on obtient les fréquences données au tableau 3.4.

Les diplotypes possibles pour le cinquième individu sont (abc, AbC) et (abC, Abc). Il nous faut donc calculer $\pi(abc \mid H^{(0)})$, $\pi(abC \mid H^{(0)})$, $\pi(Abc \mid H^{(0)})$ et $\pi(AbC \mid H^{(0)})$. Pour ce faire, il suffit d'appliquer le résultat obtenu à l'équation (3.1), ce qui équivaut à multiplier le vecteur ligne formé des fréquences relatives des haplotypes par les colonnes

Tableau 3.3 Reconstruction initiale pour Phase

	Génotype	Diplotype
1	{(a,a)(b,b)(c,c)}	(abc, abc)
2	{(a,a)(b,b)(c,c)}	(abc, abc)
3	{(a,a)(b,b)(c,c)}	(abc, abc)
4	{(a,A)(b,b)(c,C)}	(abC, Abc)
5	**{(a,A)(b,b)(c,C)}**	**(abc, AbC)** *
6	{(a,A)(b,b)(c,C)}	(abc, AbC)
7	{(a,A)(b,b)(c,C)}	(abC, Abc)
8	{(a,A)(b,B)(c,c)}	(abc, ABc)
9	{(A,A)(b,b)(C,C)}	(AbC, AbC)
10	{(A,A)(b,b)(C,C)}	(AbC, AbC)
11	{(A,A)(b,B)(c,C)}	(Abc, ABC)

*. L'individu mis à jour est identifié en caractères gras.

Tableau 3.4 Fréquence des haplotypes dans $H^{(0)}$

Haplotype	Fréquence r_α *	Fréquence relative
abc	8	0,40
abC	2	0,10
aBc	0	0,00
aBC	0	0,00
Abc	3	0,15
AbC	5	0,25
ABc	1	0,05
ABC	1	0,05
total	20	1

*. L'individu à reconstruire a été retiré.

de M correspondant aux séquences d'intérêt. Ainsi, on a

$$\begin{bmatrix} 0,4 & 0,1 & 0 & 0 & 0,15 & 0,25 & 0,05 & 0,05 \end{bmatrix} \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{matrix} \text{abc} & \text{abC} & \text{Abc} & \text{AbC} \end{matrix} \\ \begin{bmatrix} 0,9116 & 0,0277 & 0,0277 & 0,0017 \\ 0,0277 & 0,9116 & 0,0017 & 0,0277 \\ 0,0277 & 0,0017 & 0,0017 & 0,0002 \\ 0,0017 & 0,0277 & 0,0002 & 0,0017 \\ 0,0277 & 0,0017 & 0,9116 & 0,0277 \\ 0,0017 & 0,0277 & 0,0277 & 0,9116 \\ 0,0017 & 0,0002 & 0,0277 & 0,0017 \\ 0,0002 & 0,0017 & 0,0017 & 0,0277 \end{bmatrix} \end{array}$$

$$= \begin{bmatrix} 0,3721 & 0,1095 & 0,1564 & 0,2370 \end{bmatrix}.$$

On obtient alors les valeurs suivantes pour les diplotypes :

$$\begin{aligned}
\pi(abc \mid H^{(0)})\pi(AbC \mid H^{(0)}) &= 0,3721 \times 0,2370 \\
&= 0,0882; \\
\pi(abC \mid H^{(0)})\pi(Abc \mid H^{(0)}) &= 0,1095 \times 0,1564 \\
&= 0,0171.
\end{aligned}$$

Ceci nous permet de trouver les probabilités :

$$\begin{aligned}
P((abc, AbC) \mid D_{-5}^{(0)}, G) &= 0,8376; \\
P((abC, Abc) \mid D_{-5}^{(0)}, G) &= 0,1624
\end{aligned}$$

obtenues en corrigeant pour sommer à un.

On génère alors une nouvelle reconstruction pour l'individu selon la distribution précédente. On retrouve dans ce cas-ci les mêmes diplotypes que ceux donnés au tableau 3.3. On répète le processus jusqu'à atteindre la distribution stationnaire. On peut alors reporter un état de la chaîne de Markov comme estimation ponctuelle des diplotypes. Il est aussi possible de recueillir un échantillon de reconstructions, duquel on estimera les paramètres d'intérêt. Par exemple, si on s'intéresse plutôt aux fréquences des divers haplotypes dans la population, on estimera celles-ci par des moyennes empiriques obtenues sur l'échantillon de reconstructions.

3.3 Algorithme EM

3.3.1 Algorithme EM généralisé

L'algorithme EM (*Expectation Maximization*, en anglais, et Espérance-Maximisation, en français) est une méthode itérative d'estimation en deux étapes visant à maximiser une fonction de vraisemblance $L(\Psi) = P(Y \mid \Psi)$ où Y est un ensemble de données observées dit «incomplet» et Ψ un vecteur de paramètres. On utilise alors un ensemble de données complet X choisi de telle sorte que la distribution conditionnelle $P(X \mid Y, \Psi)$ est connue et que la vraisemblance complète $L_c(\Psi) = P(X \mid \Psi)$ est plus facile à

maximiser que la vraisemblance sur les données incomplètes. Notons que le choix des données complètes n'est pas unique et doit avoir pour objectif de simplifier au maximum les calculs théoriques. Le résumé qui suit reprend, à peu de choses près, les notations du livre de McLachlan et Krishnan (2008).

Chaque itération de l'algorithme se fait en deux étapes. L'étape E (Espérance) consiste à évaluer la fonction

$$W(\Psi \mid \Psi^{(k)}) = E(\log L_c(\Psi \mid Y) \mid \Psi^{(k)}),$$

où $\Psi^{(k)}$ est la valeur des paramètres à estimer après l'itération k. Notons que $\Psi^{(0)}$ est une valeur initiale prédéterminée. L'étape M (Maximisation) consiste par la suite à déterminer la valeur $\Psi^{(k+1)}$ qui maximise cette expression. Ces deux étapes sont répétées jusqu'à l'atteinte d'un critère de convergence sur les paramètres. Bien que nous n'aborderons pas la preuve ici, il a été démontré (Dempster, Laird et Rubin, 1977) que $L(\Psi^{(k+1)}) \geq L(\Psi^{(k)})$. Il est alors immédiat de constater que, si la fonction de vraisemblance sur les données observées est bornée supérieurement, l'algorithme convergera vers un maximum local de vraisemblance qui dépendra du point de départ $\Psi^{(0)}$.

Au premier regard, l'usage de l'algorithme EM généralisé ne semble pas évident, en raison de la difficulté que peut représenter l'évaluation de la fonction $W(\Psi \mid \Psi^{(k)})$. Cependant, si la distribution $P(X \mid \Psi)$ détermine une famille exponentielle de lois, l'algorithme se présente sous une forme beaucoup plus simple. Supposons que X appartient à une famille exponentielle de lois, c'est-à-dire que

$$\begin{aligned} \log(P(X \mid \Psi)) &= \sum_i [a_i(\Psi) t_i(X)] - b(\Psi) + c(X) \\ &= \mathbf{a}(\Psi)^T \mathbf{t}(X) - b(\Psi) + c(X), \end{aligned}$$

où $\mathbf{a}(\Psi)$ est une fonction vectorielle, $b(\Psi)$ et $c(X)$ des fonctions scalaires et $\mathbf{t}(X)$ un vecteur de statistiques exhaustives pour Ψ . Rappelons brièvement qu'une statistique sur un échantillon est exhaustive pour un paramètre si la connaissance de celle-ci suffit

à calculer l'estimateur à maximum de vraisemblance pour ce dernier. Dans ce cas,

$$\begin{aligned} W(\Psi \mid \Psi^{(k)}) &= E[\mathbf{a}(\Psi)^T \mathbf{t}(X) - b(\Psi) + c(X) \mid Y, \Psi^{(k)}] \\ &= \mathbf{a}(\Psi)^T E[\mathbf{t}(X) \mid Y, \Psi^{(k)}] - b(\Psi) + K_{X,\Psi^{(k)}} \end{aligned}$$

où $K_{X,\Psi^{(k)}}$ est une constante indépendante de Ψ qui n'intervient pas dans le processus de maximisation. On peut donc ignorer cette dernière dans l'évaluation de la fonction W. Pour cette raison, l'optimisation de W ne dépend de X que par la valeur moyenne des statistiques exhaustives, conditionnellement à Y et $\Psi^{(k)}$. Dans ce cas, l'étape E se résume à utiliser la distribution conditionnelle $P(X \mid \Psi, Y)$ pour évaluer l'espérance des statistiques exhaustives, c'est-à-dire

$$\mathbf{t}^{(k+1)} = E(\mathbf{t}(X) \mid Y, \Psi^{(k)}).$$

L'étape M consiste alors à trouver la valeur $\Psi^{(k+1)}$ qui maximise la vraisemblance sur les données complètes, conditionnellement aux statistiques exhaustives : $L_c(\Psi \mid \mathbf{t}^{(k+1)})$. Pour ce faire, il suffit de maximiser l'expression $\mathbf{a}(\Psi)^T \mathbf{t}^{(k+1)} - b(\Psi)$ sur Ψ par des techniques usuelles d'optimisation. En répétant ces deux étapes jusqu'à convergence, on obtient un estimateur $\hat{\Psi}$ de Ψ qui correspond à un maximum local de vraisemblance sur les données observées. Si la vraisemblance est multimodale, l'estimateur obtenu dépend du point de départ de l'algorithme. La figure 3.1 illustre les étapes successives d'un algorithme EM pour une famille exponentielle de lois. Chaque itération est suivie d'un test de convergence. Lorsque le critère est atteint, l'algorithme prend fin. Nous verrons comment cette version de l'algorithme EM permet de développer des techniques d'estimation de fréquences d'haplotypes.

3.3.2 Algorithme de Excoffier et Slatkin

Nous présentons dans cette section la méthode d'estimation des fréquences d'haplotypes dans la population décrite par Excoffier et Slatkin (1995). Pour des raisons d'homogénéité avec ce qui sera fait au chapitre suivant, la notation et le développement théorique de l'algorithme sont différents de l'article original. Ainsi, nous avons choisi de

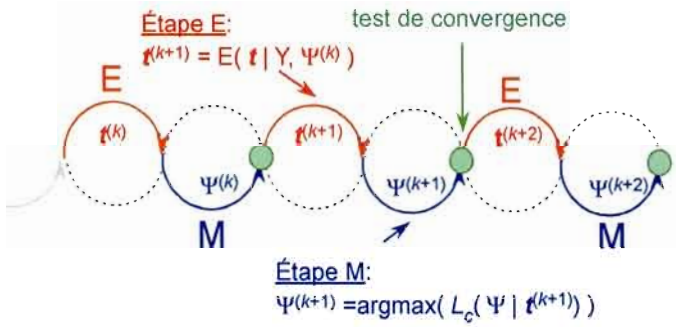

Figure 3.1 Étapes de l'algorithme EM. Les segments du haut représentent l'étape E, ceux du bas l'étape M. Chaque itération est composée d'une étape E, d'une étape M et d'un test de convergence.

retrouver celui-ci à partir de la théorie de l'algorithme EM généralisé, préparant de ce fait les bases de la nouvelle méthode présentée ultérieurement.

Considérons un échantillon aléatoire simple G de génotypes formés chacun de deux haplotypes que l'on ne peut observer directement. Supposons que l'on souhaite estimer la distribution $V(h)$ des haplotypes dans la population. Nous sommes clairement ici dans une situation de données incomplètes. Soit D l'ensemble des diplotypes parentaux des individus, c'est-à-dire pour lesquels on peut distinguer l'haplotype maternel de l'haplotype paternel. Cette précision est apportée afin d'éviter toute confusion avec le sens usuel du terme diplotype, qui signifie simplement une paire d'haplotypes, sans considération pour le statut maternel ou paternel. Cette façon de faire simplifie les calculs et notations, en nous évitant de devoir distinguer les diplotypes homozygotes et hétérozygotes. Introduisons aussi la notation $[h_i, h_j]$, représentant un diplotype parental pour lequel les séquences h_i et h_j seront respectivement l'haplotype maternel et paternel de l'individu. La notation (h_i, h_j) demeure réservée au diplotype au sens usuel. Nous noterons aussi par $[h_i, h_j] \in g$ un diplotype parental compatible avec le génotype observé g, c'est-à-dire que le génotype g peut être obtenu des deux séquences formant le diplotype. Par symétrie, on aura que $[h_i, h_j] \in g \leftrightarrow [h_j, h_i] \in g$.

Pour des raisons de calculs, considérons que G et D sont ordonnés, c'est-à-dire que les individus sont énumérés dans un ordre précis. Ceci nous évitera de devoir prendre en compte les permutations possibles sur les individus, et donc l'ajout d'une constante multinomiale qui n'intervient pas dans l'estimation. En supposant des croisements aléatoires, on remarque qu'observer les diplotypes parentaux de D équivaut à choisir chacun des haplotypes les composant de manière indépendante selon la distribution V. En effet, on peut modéliser la situation comme si chaque individu de l'échantillon choisissait une séquence maternelle puis paternelle selon la distribution V. Explicitement,

$$P(D \mid V) = \prod_{d \in D} V(h_{d,1}) V(h_{d,2}),$$

où $[h_{d,1}, h_{d,2}]$ forme le diplotype d. En regroupant les termes semblables, l'expression précédente devient :

$$P(D \mid V) = \prod_{h} V(h)^{m_h},$$

où m_h est la fréquence des séquences de type h dans D. En prenant le logarithme,

$$\log(P(D \mid V)) = \sum_{h} \log(V(h)) m_h,$$

on reconnaît la forme décrite précédemment d'une famille exponentielle de lois où $\mathbf{a}(\Psi) = \log(V)$ et les fréquences m_h sont les statistiques exhaustives. Si l'on connaissait effectivement les diplotypes, on n'aurait qu'à estimer chaque fréquence théorique $V(h)$ par la fréquence empirique m_h/m correspondante dans l'échantillon, où $m = \sum_h m_h$. Puisque les diplotypes ne sont pas disponibles, les fréquences empiriques ne sont pas observables directement. Ces dernières peuvent toutefois être estimées, si l'on suppose les paramètres $V(h)$ connus. Cette structure en boucle de l'estimation est en fait la base de l'algorithme EM, qui consiste à estimer à tour de rôle les paramètres $V(h)$ et les fréquences m_h/m, jusqu'à convergence.

Développons cette idée de manière plus rigoureuse, en se ramenant à la théorie de l'algorithme EM généralisé. Puisque la vraisemblance décrit une famille exponentielle de lois, on obtient la fonction

$$W(V \mid V^{(k)}) = \sum_{h} \log\left(V(h)\right) m_h^{(k+1)},$$

à maximiser sous la contrainte que $\sum_h V(h) = 1$. Ceci équivaut en pratique à retrouver l'estimateur à maximum de vraisemblance sur les données complètes, par les techniques usuelles. Afin de tenir compte de la contrainte, on peut incorporer un multiplicateur de Lagrange à l'équation précédente, ce qui réduit le système à l'expression

$$W_L(V \mid V^{(k)}) = \sum_h \log\left(V(h)\right) m_h^{(k+1)} + \lambda \left(1 - \sum_h V(h)\right)$$

qui peut être maximisée en annulant les dérivées partielles. Ainsi, on obtient que

$$\frac{\partial W_L}{\partial V(h)} = 0 \leftrightarrow V(h) = \frac{m_h^{(k+1)}}{\lambda}$$

et $\sum_h V(h) = 1$ implique que $\lambda = \Sigma_h m_h^{(k+1)}$. L'étape M de l'algorithme se résume par le fait même à calculer

$$V^{(k+1)}(h) = \frac{m_h^{(k+1)}}{m},$$

où $m = \Sigma_h m_h^{(k+1)}$ est le nombre, constant, d'haplotypes dans l'échantillon.

Afin de compléter l'algorithme, il reste donc à déterminer la distribution conditionnelle $P(D \mid G, V^{(k)})$ nous permettant de calculer $m_h^{(k+1)} = E_{V^k}(m_h \mid G)$ pour tout haplotype h. Nous avons vu précédemment que la probabilité d'un diplotype parental est donnée par le produit des probabilités marginales de ses haplotypes. En considérant tous les diplotypes parentaux ayant pu donner lieu au génotype observé, on peut en déduire la probabilité conditionnelle d'un diplotype parental $[h_i, h_j] \in g$:

$$\begin{aligned} P([h_i, h_j] \mid V) &= V(h_i)V(h_j), \\ P(g \mid V) &= \sum_{[h_l, h_s] \in g} V(h_l)V(h_s), \\ P([h_i, h_j] \mid g, V) &= \frac{V(h_i)V(h_j)}{\sum_{[h_l, h_s] \in g} V(h_l)V(h_s)}. \end{aligned} \tag{3.2}$$

Cette probabilité conditionnelle nous permet de déterminer le nombre moyen de séquences de type h portées par un échantillon d'individus présentant un génotype compatible g. Parmi n_g individus arborant ce génotype, il y en aura en moyenne $n_g P([h, *] \mid g, V)$ qui porteront la séquence h sur l'haplotype maternel, et $n_g P([*, h] \mid g, V)$ qui la porteront sur l'haplotype paternel. La probabilité de porter la séquence h sur l'haplotype maternel

est obtenue en sommant sur tous les haplotypes paternels compatibles, et vice versa. Notons que, s'il n'y a pas d'informations manquantes sur les génotypes, il existe un unique haplotype h_g tel que $[h, h_g] \in g$. Dans ce cas,

$$\begin{aligned} P([h,*] \mid g, V) &= P([h,h_g] \mid g, V), \\ P([*,h] \mid g, V) &= P([h_g,h] \mid g, V). \end{aligned}$$

Les deux probabilités précédentes étant égales par symétrie, le nombre moyen de copies de h parmi n_g individus arborant le génotype g est de $2n_g P([h,*] \mid g, V)$. On obtient alors pour l'ensemble de l'échantillon,

$$m_h = \sum_{g \in G} 2n_g P([h,*] \mid g, V).$$

On évalue enfin $m_h^{(k+1)}$ en remplaçant les paramètres $V(h)$ par les valeurs $V^{(k)}(h)$.

En résumé, l'algorithme EM décrit par Excoffier et Slatkin consiste dans un premier temps à déterminer une distribution initiale $V^{(0)}$. Une distribution uniforme est généralement appropriée. Par la suite, on effectue les étapes E et M de l'algorithme jusqu'à ce que la différence entre $V^{(k+1)}$ et $V^{(k)}$ atteigne un certain critère de convergence. À l'étape E, on obtient $m^{(k)}$ de $V^{(k)}$ en évaluant, pour toute séquence h,

$$m_h^{(k+1)} = \sum_{g \in G} 2n_g P([h,*] \mid g, V^{(k)}).$$

Pour ce qui est de l'étape M, on déduit $V^{(k+1)}$ de $m^{(k+1)}$ en calculant

$$V^{(k+1)}(h) = m_h^{(k+1)}/m.$$

La figure 3.2 illustre ces étapes.

Une fois la distribution V estimée, il est possible, par exemple, d'en déduire un estimateur des diplotypes des individus en utilisant la distribution conditionnelle décrite à l'équation (3.2). En effet, on peut estimer celle-ci en remplaçant V par son estimé $\hat{V}$.

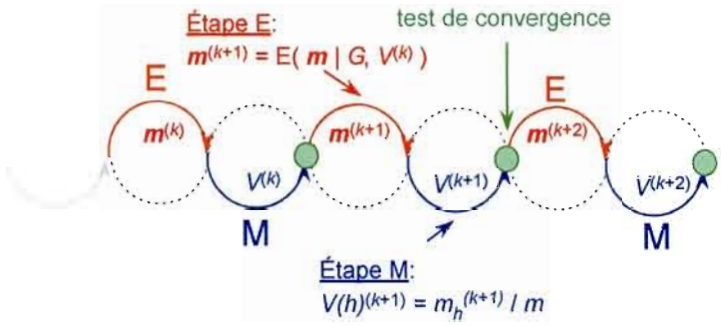

Figure 3.2 Étapes de l'algorithme de Excoffier et Slatkin. Les segments du haut représentent l'étape E, ceux du bas l'étape M. Chaque itération est composée d'une étape E, d'une étape M et d'un test de convergence.

3.3.3 Illustration

Reprenons de nouveau l'exemple donné au tableau 3.1, page 37. Puisqu'on ne dispose d'aucune information préalable sur les haplotypes, on prend comme point de départ une distribution uniforme. Puisqu'on a trois marqueurs binaires, le nombre total d'haplotypes théoriques est de $2^3 = 8$. Ainsi, pour chaque haplotype h, on a $V(h)^{(0)} = 1/8$. Considérons le génotype $\{(a, A)(b, b)(c, C)\}$, dont la fréquence dans l'échantillon est de quatre. Les diplotypes parentaux compatibles sont $[abc, AbC]$, $[AbC, abc]$, $[abC, Abc]$ et $[Abc, abC]$. On calcule aisément les probabilités associées à ces diplotypes par le produit des probabilités marginales des haplotypes.

$$\begin{aligned} P([abc, AbC] \mid V^{(0)}) &= V^{(0)}(abc)V^{(0)}(AbC) \\ &= 1/64, \end{aligned}$$

$$\begin{aligned} P([AbC, abc] \mid V^{(0)}) &= V^{(0)}(AbC)V^{(0)}(abc) \\ &= 1/64, \end{aligned}$$

$$\begin{aligned} P([abC, Abc] \mid V^{(0)}) &= V^{(0)}(abC)V^{(0)}(Abc) \\ &= 1/64, \end{aligned}$$

$$\begin{aligned} P([Abc, abC] \mid V^{(0)}) &= V^{(0)}(Abc)V^{(0)}(abC) \\ &= 1/64. \end{aligned}$$

En sommant ces résultats, on trouve pour ce génotype :

$$P(\{(a,A)(b,b)(c,C)\} \mid V^{(0)}) = 1/16.$$

On en déduit les probabilités conditionnelles ;

$$\begin{aligned} P([abc, AbC] \mid \{(a,A)(b,b)(c,C)\}, V^{(0)}) &= 0,25 \\ P([AbC, abc] \mid \{(a,A)(b,b)(c,C)\}, V^{(0)}) &= 0,25 \\ P([abC, Abc] \mid \{(a,A)(b,b)(c,C)\}, V^{(0)}) &= 0,25 \\ P([Abc, abC] \mid \{(a,A)(b,b)(c,C)\}, V^{(0)}) &= 0,25. \end{aligned}$$

La seconde ligne du tableau 3.5 est obtenue en multipliant le résultat précédent par huit, soit deux fois la fréquence de ce génotype, ce qui correspond au nombre de séquences haploïdes qui y sont associées. On construit l'ensemble du tableau en répétant l'opération pour tous les génotypes. En sommant sur chaque colonne de celui-ci, on calcule aisément les fréquences $m_h^{(1)}$. De celles-ci, on obtient directement $V^{(1)}$ en divisant par $m = 22$, le nombre total de séquences. On répète le processus jusqu'à convergence. Le tableau 3.6 donne les fréquences moyennes pour la seconde itération, tandis que le tableau 3.7 donne l'évolution de la distribution après deux itérations.

Tableau 3.5 Première itération de l'algorithme EM

génotype (g) *	n_g	abc	abC	aBc	aBC	Abc	AbC	ABc	ABC
$\{(a,a)(b,b)(c,c)\}$	3	6,0	0	0	0	0	0	0	0
$\{(\mathbf{a},\mathbf{A})(\mathbf{b},\mathbf{b})(\mathbf{c},\mathbf{C})\}$	4	2,0	2,0	0	0	2,0	2,0	0	0
$\{(a,A)(b,B)(c,c)\}$	1	0,5	0	0,5	0	0,5	0	0,5	0
$\{(A,A)(b,b)(C,C)\}$	2	0	0	0	0	0	4,0	0	0
$\{(A,A)(b,B)(c,C)\}$	1	0	0	0	0	0,5	0,5	0,5	0,5
$m_h^{(1)}$		8,5	2,0	0,5	0	3,0	6,5	1,0	0,5

*. Chaque ligne représente un génotype g. Sur chacune, on retrouve le nombre moyen de copies des divers haplotypes portés par n_g individus présentant le génotype g.

Tableau 3.6 Seconde itération de l'algorithme EM

génotype (g) *	n_g	abc	abC	aBc	aBC	Abc	AbC	ABc	ABC
$\{(a,a)(b,b)(c,c)\}$	3	6,00	0	0	0	0	0	0	0
$\{(a,A)(b,b)(c,C)\}$	4	3,61	0,39	0	0	0,39	3,61	0	0
$\{(a,A)(b,B)(c,c)\}$	1	0,85	0	0,15	0	0,15	0	0,85	0
$\{(A,A)(b,b)(C,C)\}$	2	0	0	0	0	0	4,00	0	0
$\{(A,A)(b,B)(c,C)\}$	1	0	0	0	0	0,19	0,81	0,81	0,19
$m_h^{(2)}$		10,46	0,39	0,15	0,00	0,73	8,42	1,66	0,19

*. Chaque ligne représente un génotype g. Sur chacune, on retrouve le nombre moyen de copies des divers haplotypes portés par n_g individus présentant le génotype g.

Tableau 3.7 Évolution de la distribution $V(h)$

h	abc	abC	aBc	aBC	Abc	AbC	ABc	ABC
$V(h)^{(0)}$	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
$V(h)^{(1)}$	0,3864	0,0909	0,0227	0,0000	0,1364	0,2955	0,0455	0,0227
$V(h)^{(2)}$	0,4755	0,0178	0,0068	0,0000	0,0332	0,3827	0,0754	0,0086

3.4 Performances et comparaisons

3.4.1 Performances

Nous avons présenté dans les sections précédentes des algorithmes permettant d'estimer les haplotypes. La comparaison de ces méthodes a fait l'objet de plusieurs publications. Une des études les plus complètes sur ce sujet a été effectuée pour le compte du projet internationnal HapMap (Marchini *et al.*, 2006). Le projet international HapMap vise à répertorier les variations dans le code génétique humain. Celui-ci regroupe des chercheurs du Canada, de la Chine, du Japon, du Nigeria, du Royaume-Uni et des États-Unis. Les auteurs de l'étude ont conclu que la méthode Phase était la plus performante parmi celles comparées. Notons que l'algorithme de parcimonie n'en faisait pas partie. Par performante, nous entendons ici que les taux d'erreurs sur les estimations étaient les plus faibles. Ceci dit, les méthodes basées sur l'algorithme EM ont obtenu des résultats comparables à Phase sur des données simulées selon le principe de coalescence.

Les résultats obtenus par les chercheurs du projet HapMap confirment ceux reportés dans d'autres publications (Xu et al., 2004; Stephens et Scheet, 2005; Huang, Chao et Ting, 2005; Qin, Niu et Liu, 2002). Étant donné que les résultats de chaque comparaison dépendent grandement des hypothèses retenues et des statistiques utilisées, il nous apparaît inadéquat d'en reporter les résultats numériques hors de leur contexte. Ceci dit, les conclusions générales de ces articles sont similaires : pour des données simulées selon le principe de coalescence, la méthode Phase a donné de meilleurs résultats dans la grande majorité des comparaisons. Par contre, il semblerait que cet avantage disparaît lorsque des données réelles sont utilisées (Qin, Niu et Liu, 2002). Ainsi, la comparaison des diverses méthodes ne permet pas d'en déclarer une nettement supérieure.

Quelle que soit la méthode utilisée, l'estimation des diplotypes comporte généralement un grand nombre d'erreurs. En effet, l'espace sur les haplotypes croît de façon exponentielle avec le nombre de marqueurs. Ainsi, la probabilité d'obtenir les diplotypes de tous les individus avec exactitude tend rapidement vers zéro. Pour cette raison, les auteurs

des différentes études citées utilisent des taux d'erreurs plus réalistes, en comparant, par exemple, les fréquences théoriques avec celles estimées. En effet, il est généralement plus aisé d'estimer les fréquences relatives des séquences haploïdes que de retrouver les diplotypes formant les génotypes. Or, les erreurs sur les diplotypes peuvent avoir une incidence importante sur les méthodes de cartographie qui en dépendent. La façon d'intégrer l'estimation des haplotypes dans ces méthodes prend donc toute son importance.

3.4.2 Estimation des haplotypes en cartographie

Plusieurs méthodes de cartographie génétique ont été confrontées à la problématique de l'estimation des haplotypes. Pour certaines d'entre elles, le modèle et l'algorithme utilisés permettent l'estimation simultanée des haplotypes et de la position de la mutation. Ainsi, en imposant des contraintes sur les événements, il est possible de reconstruire les diplotypes probables des individus par la construction d'un graphe de recombinaison ancestral. Cette stratégie a été utilisée dans la méthode Margarita (Minichiello et Durbin, 2006). Il est à noter que, dans ce cas, la distribution développée par les auteurs ne correspond pas explicitement à un modèle établi, mais plutôt à des principes généraux et critères choisis par eux, dans le cadre du principe de coalescence. Dans le contexte d'un modèle d'échantillonnage pondéré, cette technique ne peut être utilisée, puisque la distribution proposée n'est pas connue explicitement.

Certains algorithmes basés sur des chaînes de Markov de Monte Carlo permettent de considérer les diplotypes comme des variables latentes. Dans ce cas, ceux-ci sont actualisés à chaque étape de la chaîne, en même temps que les paramètres estimés (Morris, Whittaker et Balding, 2000). Selon les auteurs, cette approche simultanée est préférable à une estimation ponctuelle préalable des haplotypes par une méthode externe telle Phase. Selon eux, les erreurs sur les haplotypes auraient un impact non négligeable sur les résultats obtenus, en plus d'exagérer le déséquilibre de liaison. Enfin, ceux-

ci mentionnent que les techniques actuelles d'estimation ne tiennent pas compte des phénotypes, ce qui leur paraît inapproprié dans le contexte d'une étude de type cas-témoins.

La méthode DHSMAP (McPeek et Strahs, 1999) utilise quant à elle une chaîne de Markov cachée sur le statut ancestral des marqueurs pour calculer la vraisemblance des génotypes de l'échantillon. Les étapes de la chaîne correspondent alors aux marqueurs de la séquence, dans l'ordre. Selon leur modèle, l'allèle à un marqueur donné dépend uniquement du marqueur précédent et de leur état respectif (ancestral ou non). Ainsi, les marqueurs sont considérés par paires superposées, de manière analogue aux fenêtres de la figure 2.3, page 28. Dans ce contexte, en supposant des marqueurs binaires, il existe seulement $2^2 = 4$ haplotypes possibles. À chaque génotype correspond donc au plus deux diplotypes compatibles. Sans entrer dans les détails, les fréquences d'haplotypes pour chaque paire de marqueurs sont estimées en même temps que les autres paramètres de la méthode par un algorithme EM généralisé. Enfin, sans estimer explicitement les diplotypes, chaque configuration possible est alors considérée dans le calcul de la vraisemblance.

Tandis que certains algorithmes de cartographie intègrent la problématique des génotypes non résolus, d'autres nécessitent un travail en deux étapes. Ainsi, la méthode TreeLD (Zöllner et Pritchard, 2005), basée sur des arbres de coalescence, exige un échantillon de diplotypes. Pour en obtenir un, les auteurs suggèrent d'utiliser une estimation ponctuelle donnée par un algorithme d'estimation des haplotypes, tel Phase. Cette approche n'est pas à rejeter, car elle peut donner des résultats intéressant si l'estimation préalable ne comporte pas trop d'erreurs. Elle nous paraît cependant risquée, puisqu'elle ignore l'incertitude liée à l'estimation des haplotypes. Nous partageons en ce sens les réserves relevées dans Morris, Whittaker et Balding (2000).

3.4.3 Choix d'une méthode

Tous les algorithmes d'estimation des haplotypes que nous avons présentés ont une même lacune, soit celle d'être difficilement applicable à des séquences comprenant un grand nombre de marqueurs. En effet, l'espace sur les haplotypes possibles devient alors très grand. Par conséquent, on se retrouve avec des problèmes de mémoire ou une extrême lenteur des programmes informatiques. Une solution à cette problématique commune consiste à traiter de manière indépendante des segments de la séquence d'origine, puis de les combiner pour former un haplotype entier. Ainsi, dans le cas de l'algorithme EM, on effectue dans un premier temps une partition de la séquence en segments de taille jugée raisonnable. Pour chacun, on utilise l'algorithme pour estimer les fréquences d'haplotypes. Dans une seconde étape, on traite chaque segment comme un marqueur dont les allèles multiples sont donnés par les haplotypes les plus fréquents. On applique finalement l'algorithme EM sur ces nouveaux marqueurs pour obtenir l'estimation globale (Qin, Niu et Liu, 2002). Les résultats obtenus n'optimisent pas la vraisemblance exacte, mais plutôt une pseudo vraisemblance dépendante du découpage, qui n'est pas sans rappeler la vraisemblance composite définie au chapitre précédent. La partition de la séquence est une solution utilisée par plusieurs méthodes d'estimation des haplotypes, dont la version actuelle de Phase (Stephens et Scheet, 2005) et d'autres méthodes bayésiennes (Niu et al., 2002). Dans le contexte de la méthode MapArg, cette partition sera traitée par l'utilisation de la vraisemblance composite.

Les méthodes décrites présentent, à nos yeux, une autre lacune commune. Dans le cas d'études de type cas-témoins, on s'attend à ce que la connaissance du phénotype des individus apporte une certaine information sur les diplotypes. De plus, la méthode de cartographie que nous tentons de généraliser, MapArg, nécessite la connaissance des allèles au gène causal. Or, aucune des méthodes présentées ne tient compte des phénotypes. D'ailleurs, nos recherches n'ont pas permis d'en trouver une qui estime explicitement les allèles au gène causal. Il serait toutefois possible d'obtenir un algorithme EM et

une nouvelle version de Phase conditionnels aux phénotypes et permettant d'obtenir l'information recherchée.

Puisque la méthode de cartographie que nous tentons de généraliser est basée sur un modèle d'échantillonnage pondéré, il nous apparaît souhaitable que l'estimation des haplotypes soit associée à une distribution de probabilité connue. Dans ce contexte, l'algorithme EM nous semble tout indiqué. En effet, une fois les paramètres estimés, il est possible de générer un grand nombre de reconstructions pour les diplotypes, en associant une probabilité à chacune. De son côté, Phase permet de générer un grand nombre de diplotypes, mais sans probabilité associée. Comme il est basé sur un échantillonnage de Gibbs, les estimations obtenues pour les paramètres sont aléatoires. En effet, deux applications successives ne permettront pas d'obtenir les mêmes résultats pour les estimateurs. L'algorithme EM a aussi l'avantage d'être plus aisé à modifier et à programmer. Enfin, la rapidité d'exécution de l'algorithme EM le rend moins lourd à utiliser que Phase. Par conséquent, nous développerons au prochain chapitre une variante de l'algorithme EM qui prendra en considération les allèles au gène causal et les phénotypes des individus.

CHAPITRE IV

ALGORITHME EM CONDITIONNEL AUX PHÉNOTYPES

Nous avons présenté au précédent chapitre quelques méthodes d'estimation des haplotypes. Malheureusement, aucune ne peut être appliquée directement à notre problème. En effet, telle que la méthode MapArg est construite, il nous faut estimer le statut au gène causal, en plus des haplotypes, ce qu'aucune méthode ne permet de faire actuellement. Or, s'il existe un déséquilibre de liaison entre les marqueurs de référence et le gène causal, la distribution des allèles à ces marqueurs ne sera pas la même parmi les haplotypes porteurs et non porteurs. Ce déséquilibre est d'ailleurs à l'origine des méthodes statistiques de cartographie génétique. Pour cette raison, nous considérerons deux distributions pour les haplotypes dans la population. Puisqu'il existe une dépendance certaine entre le phénotype et le statut au gène causal, nous estimerons les paramètres de ces distributions par un algorithme EM conditionnel aux phénotypes. Nous verrons comment cette nouvelle méthode est robuste face à un échantillonnage à proportion fixée de cas. Enfin, nous en déduirons une technique d'estimation du modèle de pénétrance. Notons qu'un résumé des notations utilisées dans ce chapitre est présenté à l'annexe C.

4.1 Estimation des distributions haploïdes

4.1.1 Vraisemblance complète et étape M

Nous supposerons ici une population très large d'individus diploïdes, de telle sorte que les fréquences théoriques et observées soient les mêmes (loi des grands nombres). Cette po-

pulation sera aussi supposée en équilibre de Hardy-Weinberg avec croisements aléatoires. L'équilibre de Hardy-Weinberg est une hypothèse selon laquelle la distribution des allèles est constante de générations en générations. Celle-ci est généralement vérifiée dans le cas de populations qui n'ont pas connu de variations importantes de leur patrimoine génétique depuis plusieurs générations. Rappelons que nous considérons ici un caractère de type dichotomique, c'est-à-dire cas-témoins pour lequel un seul gène causal, ou mutation, est impliqué sur la région chromosomique concernée.

Notons par V_0 la distribution des haplotypes non porteurs et V_1 celle des porteurs. Ainsi, $V_0(h)$ sera la proportion de séquences haploïdes de type h parmi toutes celles qui sont non porteuses du gène causal. On associera à celui-ci un modèle de pénétrance $F = (f_0, f_1, f_2)$ où f_i est la probabilité pour un individu de présenter le caractère, sachant qu'il porte i copies de la mutation causale. Soit T le statut d'un individu pour la mutation, de sorte que $T = 00$, $T = 01$, $T = 10$ et $T = 11$ représentent respectivement un individu non porteur, un haplotype paternel mutant, un haplotype maternel mutant et un doublement porteur. Enfin, notons par p la proportion de séquences haploïdes porteuses dans la population et par f la proportion d'individus diploïdes présentant le caractère qui nous préoccupe.

Puisque la population est supposée en équilibre de Hardy-Weinberg, on peut aisément déterminer la distribution des allèles dans la population. Ainsi, la probabilité qu'un individu soit doublement porteur est de p^2 et celui-ci présentera le caractère d'intérêt avec probabilité f_2, ce qui nous donne une probabilité conjointe de $f_2 p^2$. Chacune des probabilités regroupées au tableau 4.1 est obtenue par un raisonnement similaire.

Tableau 4.1 Distribution des allèles au gène causal dans la population

Nombre de copies du gène causal	Cas	Témoin	total
T=00	$f_0(1-p)^2$	$(1-f_0)(1-p)^2$	$(1-p)^2$
T=01	$f_1p(1-p)$	$(1-f_1)p(1-p)$	$p(1-p)$
T=10	$f_1p(1-p)$	$(1-f_1)p(1-p)$	$p(1-p)$
T=11	f_2p^2	$(1-f_2)p^2$	p^2
total	f	$(1-f)$	1

Soit G, un échantillon aléatoire simple de génotypes tirés de la population que nous venons de décrire, auxquels sont associés les phénotypes observés Φ. Notre objectif étant dans un premier temps d'estimer les distributions V_0 et V_1, nous nous retrouvons dans une situation de données incomplètes. Considérons donc l'ensemble de données complètes composé des phénotypes Φ et des diplotypes parentaux D des individus, incluant le gène causal. Nous utiliserons la notation h^0 pour représenter un haplotype non porteur de type h sur les marqueurs de référence et h^1 pour un porteur. Une fois connu le statut au gène causal, l'identité des séquences est uniquement déterminée par les distributions V_0 et V_1. Les deux séquences qui composent le diplotype sont alors choisies de manière indépendante dans leur distribution respective. La probabilité du diplotype parental $[h_i^{\delta_1}, h_j^{\delta_2}]$ est alors donnée par

$$P([h_i^{\delta_1}, h_j^{\delta_2}] \mid V_0, V_1) = V_{\delta_1}(h_i)V_{\delta_2}(h_j)P(T = \delta_1\delta_2),$$

où $\delta_1, \delta_2 \in \{0, 1\}$. Puisque le phénotype ne dépend du diplotype que par le gène causal, on en déduit la probabilité conjointe du diplotype $[h_i^{\delta_1}, h_j^{\delta_2}]$ et du phénotype ϕ :

$$\begin{aligned} P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0, V_1) &= P(\phi \mid [h_i^{\delta_1}, h_j^{\delta_2}], V_0, V_1)P([h_i^{\delta_1}, h_j^{\delta_2}] \mid V_0, V_1) \\ &= P(\phi \mid T = \delta_1\delta_2)P(T = \delta_1\delta_2)V_{\delta_1}(h_i)V_{\delta_2}(h_j) \\ &= P(\phi, T = \delta_1\delta_2)V_{\delta_1}(h_i)V_{\delta_2}(h_j). \end{aligned} \tag{4.1}$$

Les probabilités $P(\phi, T = \delta_1\delta_2)$ étant données dans le tableau 4.1, on obtient explicitement :

$$\begin{aligned}
P([h_i^0, h_j^0], \phi = 1, \mid V_0, V_1) &= f_0(1-p)^2 V_0(h_i)V_0(h_j), \\
P([h_i^0, h_j^1], \phi = 1 \mid V_0, V_1) &= f_1 p(1-p) V_0(h_i)V_1(h_j), \\
P([h_i^1, h_j^0], \phi = 1 \mid V_0, V_1) &= f_1 p(1-p) V_1(h_i)V_0(h_j), \\
P([h_i^1, h_j^1], \phi = 1 \mid V_0, V_1) &= f_2 p^2 V_1(h_i)V_1(h_j), \\
P([h_i^0, h_j^0], \phi = 0 \mid V_0, V_1) &= (1-f_0)(1-p)^2 V_0(h_i)V_0(h_j), \\
P([h_i^0, h_j^1], \phi = 0 \mid V_0, V_1) &= (1-f_1) p(1-p) V_0(h_i)V_1(h_j), \\
P([h_i^1, h_j^0], \phi = 0 \mid V_0, V_1) &= (1-f_1) p(1-p) V_1(h_i)V_0(h_j), \\
P([h_i^1, h_j^1], \phi = 0 \mid V_0, V_1) &= (1-f_2) p^2 V_1(h_i)V_1(h_j).
\end{aligned}$$

Puisque les individus sont indépendants entre eux, la probabilité conjointe est obtenue par le produit des probabilités marginales. On obtient la vraisemblance sur les données complètes :

$$\begin{aligned}
L_c(V_0, V_1) &= P(D, \Phi \mid V_0, V_1) \\
&= \prod_i P(d_i, \phi_i \mid V_0, V_1) \\
&= \prod_i P([h_{i1}^{\delta_{i1}}, h_{i2}^{\delta_{i2}}], \phi_i \mid V_0, V_1) \\
&= \prod_i P(\phi_i, T = \delta_{i1}\delta_{i2}) V_{\delta_{i1}}(h_{i1}) V_{\delta_{i2}}(h_{i2}),
\end{aligned}$$

où $d_i = [h_{i1}^{\delta_{i1}}, h_{i2}^{\delta_{i2}}]$ est le diplotype parental associé à l'individu i. Puisque les probabilités $P(\phi_i, T = \delta_{i1}\delta_{i2})$ ne dépendent que du modèle de pénétrance et pas des distributions à estimer, on peut écrire l'expression précédente sous la forme :

$$L_c(V_0, V_1) = K(F, p) \prod_i V_{\delta_{i1}}(h_{i1}) V_{\delta_{i2}}(h_{i2}).$$

En regroupant les termes, on peut exprimer cette vraisemblance en fonction des fréquences d'haplotypes,

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

où m_{h^0} et m_{h^1} sont respectivement le nombre de séquences non porteuses et porteuses de type h dans D. La vraisemblance obtenue correspond de nouveau à une famille exponentielle de lois, les statistiques exhaustives pour V_0 et V_1 étant les fréquences m_h. Si l'on connaissait effectivement les diplotypes, il nous suffirait donc d'estimer les fréquences théoriques par les fréquences empiriques. Or, ces diplotypes ne sont pas observables, mais pourraient être estimés si les distributions V_0 et V_1 étaient connues.

Ramenons-nous à la théorie de l'algorithme EM généralisé. Notons l'espérance des statistiques exhaustives sous la forme :

$$m_{h^\delta}^{(k+1)} = E(m_{h^\delta} \mid V_0^{(k)}, V_1^{(k)}, G, \Phi).$$

La fonction à maximiser pour construire l'algorithme EM est alors :

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

sous contraintes que $\sum_h V_0(h) = 1$ et $\sum_h V_1(h) = 1$. En incorporant un multiplicateur de Lagrange pour chaque contrainte, ce système se réduit à optimiser l'expression linéaire

$$\begin{aligned} W_L(V_0, V_1 \mid V_0^{(k)}, V_1^{(k)}) &= \sum_h \{ m_{h^0}^{(k+1)} \log(V_0(h)) + m_{h^1}^{(k+1)} \log(V_1(h)) \} \\ &\quad + \lambda_0 (1 - \sum_h V_0(h)) + \lambda_1 (1 - \sum_h V_1(h)), \end{aligned}$$

c'est-à-dire à retrouver les estimateurs à maximum de vraisemblance sur les données complètes. En annulant les dérivées partielles, on retrouve que W_L est maximale si $V_0(h) = m_{h^0}^{(k+1)}/\lambda_0$ et $V_1(h) = m_{h^1}^{(k+1)}/\lambda_1$. En appliquant les contraintes, l'étape M de l'algorithme se résume à évaluer

$$V_0(h)^{(k+1)} = \frac{m_{h^0}^{(k+1)}}{m_0^{(k+1)}}, \tag{4.2}$$

$$V_1(h)^{(k+1)} = \frac{m_{h^1}^{(k+1)}}{m_1^{(k+1)}}, \tag{4.3}$$

où $m_0^{(k+1)} = \sum_h m_{h^0}^{(k+1)}$ et $m_1^{(k+1)} = \sum_h m_{h^1}^{(k+1)}$ représentent le nombre moyen de séquences non porteuses et porteuses après l'itération k. Notons qu'il est possible que

$m_0^{(k+1)}$ ou $m_1^{(k+1)}$ converge vers zéro, ce qui correspondrait au cas où toutes les séquences présenteraient le même allèle au gène causal. Dans ce cas, on ne disposerait d'aucune information concernant la distribution correspondante que l'on supposera uniforme.

4.1.2 Espérances conditionnelles et étape E

Nous avons vu comment calculer les distributions $V_0^{(k+1)}$ et $V_1^{(k+1)}$, sachant la valeur moyenne des fréquences d'haplotypes. Afin de compléter l'algorithme, il est maintenant nécessaire d'évaluer les espérances conditionnelles $m_{h^\delta}^{(k+1)} = E(m_{h^\delta} \mid V_0^{(k)}, V_1^{(k)}, G, \Phi)$. Nous avons donné à l'équation (4.1) la probabilité d'un diplotype parental et d'un phénotype donné. Exprimons maintenant cette même probabilité sous une forme qui permettra des simplifications et une généralisation de la méthode par la suite.

Nous avions

$$P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0, V_1) = P(\phi, T = \delta_1\delta_2) V_{\delta_1}(h_i) V_{\delta_2}(h_j).$$

En conditionnant selon le phénotype,

$$P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0, V_1) = P(\phi) P(T = \delta_1\delta_2 \mid \phi) V_{\delta_1}(h_i) V_{\delta_2}(h_j), \tag{4.4}$$

où $P(\phi = 1) = f$ et $P(\phi = 0) = 1 - f$. Les probabilités conditionnelles aux phénotypes peuvent être déduites aisément du tableau 4.1 :

$$P(T = 00 \mid \phi = 1) = \frac{f_0(1-p)^2}{f}, \tag{4.5}$$

$$P(T = 01 \mid \phi = 1) = \frac{f_1 p(1-p)}{f}, \tag{4.6}$$

$$P(T = 10 \mid \phi = 1) = \frac{f_1 p(1-p)}{f}, \tag{4.7}$$

$$P(T = 11 \mid \phi = 1) = \frac{f_2 p^2}{f}, \tag{4.8}$$

$$P(T = 00 \mid \phi = 0) = \frac{(1-f_0)(1-p)^2}{1-f}, \tag{4.9}$$

$$P(T = 01 \mid \phi = 0) = \frac{(1-f_1)p(1-p)}{1-f}, \tag{4.10}$$

$$P(T = 10 \mid \phi = 0) = \frac{(1-f_1)p(1-p)}{1-f}, \tag{4.11}$$

$$P(T = 11 \mid \phi = 0) = \frac{(1-f_2)p^2}{1-f}. \tag{4.12}$$

On déduit la probabilité conjointe du génotype g et du phénotype ϕ en sommant sur tous les diplotypes parentaux compatibles :

$$\begin{aligned} P(g, \phi \mid V_0, V_1) &= \sum_{[h_i^{\delta_1}, h_j^{\delta_2}] \in g} P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0, V_1) \\ &= P(\phi) \sum_{[h_i^{\delta_1}, h_j^{\delta_2}] \in g} P(T = \delta_1\delta_2 \mid \phi) V_{\delta_1}(h_i) V_{\delta_2}(h_j). \end{aligned} \tag{4.13}$$

Enfin, la probabilité d'un diplotype, conditionnelle au phénotype et au génotype, est obtenue par le quotient des équations (4.4) et (4.13), où $P(\phi)$ s'annule au numérateur et au dénominateur :

$$P([h_i^{\delta_1}, h_j^{\delta_2}] \mid g, \phi, V_0, V_1) = \frac{P(T = \delta_1\delta_2 \mid \phi) V_{\delta_1}(h_i) V_{\delta_2}(h_j)}{\sum_{[h_i^{\beta_1}, h_j^{\beta_2}] \in g} P(T = \beta_1\beta_2 \mid \phi) V_{\beta_1}(h_i) V_{\beta_2}(h_j)}. \tag{4.14}$$

Ainsi, la probabilité conditionnelle ne dépend que des distributions V_0, V_1 et des probabilités $P(T \mid \phi)$. Nous verrons dans une prochaine section comment cette caractéristique permet d'assouplir les hypothèses sur le mode d'échantillonnage.

Évaluons maintenant l'espérance conditionnelle $m_{h\delta}^{(k+1)}$. Considérons $n_{g,\phi}$ individus portant le génotype g et le phénotype ϕ. De ces individus, $n_{g,\phi} P([h^\delta, *] \mid g, \phi, V_0, V_1)$ auront reçu la séquence h^δ de leur mère et $n_{g,\phi} P([*, h^\delta] \mid g, \phi, V_0, V_1)$ de leur père. Comme précédemment, la probabilité conditionnelle de présenter une certaine séquence sur l'haplotype maternel est obtenue en sommant sur tous les haplotypes paternels compatibles, et vice versa. Rappelons que, s'il n'y a pas d'informations manquantes sur les génotypes,

il y a une unique séquence h_g telle que $[h, h_g] \in g$. Dans ce cas,

$$\begin{aligned} P([h^\delta, *] \mid g, \phi, V_0, V_1) &= P([h^\delta, h_g^0] \mid g, \phi, V_0, V_1) \\ &\quad + P([h^\delta, h_g^1] \mid g, \phi, V_0, V_1), \\ P([*, h^\delta] \mid g, \phi, V_0, V_1) &= P([h_g^0, h^\delta] \mid g, \phi, V_0, V_1) \\ &\quad + P([h_g^1, h^\delta] \mid g, \phi, V_0, V_1). \end{aligned}$$

Ces deux probabilités étant égales par symétrie, le nombre moyen de copies de h^δ portées par $n_{g,\phi}$ individus présentant ce profil est de $2n_{g,\phi}P([h^\delta, *] \mid g, \phi, V_0, V_1)$. On retrouve $m_{h_\delta}^{(k+1)}$ en sommant sur tous les génotypes et phénotypes. L'étape E de l'algorithme revient alors à évaluer, pour tout h^δ,

$$m_{h^\delta}^{(k+1)} = \sum_{(g,\phi)\in(G,\Phi)} 2n_{g,\phi}P([h^\delta, *] \mid g, \phi, V_0^{(k)}, V_1^{(k)}). \tag{4.15}$$

4.1.3 Échantillonnage à proportion déterminée de cas

Jusqu'à maintenant, nous avons supposé que l'on disposait d'échantillons aléatoires simples d'individus de la population. Toutefois, les études de type cas-témoins se basent rarement sur ce type d'échantillon. En effet, les caractères étudiés étant rares, un échantillon aléatoire simple risquerait de ne contenir aucun individu cas. Pour cette raison, ces études se basent habituellement sur des échantillons pour lesquels la proportion d'individus cas est fixée, c'est-à-dire que l'on a choisi de manière aléatoire un nombre n_1 d'individus parmi les cas dans la population, et $n - n_1$ parmi les témoins. La proportion de cas dans ce type d'échantillon est alors donnée par $\omega = n_1/n$. Dans cette situation, les algorithmes EM traditionnels ne sont pas appropriés. Cette limite se pose d'ailleurs pour la plupart des méthodes d'estimation des haplotypes. En effet, cette pratique modifie la distribution des allèles. Ainsi, les proportions données au tableau 4.1, de même que le modèle de pénétrance, ne reflètent plus ce qui est attendu dans l'échantillon. Toutefois, nous allons démontrer que l'algorithme EM que nous venons de décrire est robuste face à ce type d'échantillonnage.

Dans un premier temps, illustrons comment les proportions sont affectées. Puisque les individus cas et témoins ont été choisis de manière aléatoire simple dans leur population respective, les probabilités conditionnelles aux phénotypes demeurent invariantes. Ainsi, les probabilités $P(T \mid \phi, n_1) = P(T \mid \phi)$ sont celles données aux équations (4.5) à (4.12). On peut alors établir une distribution attendue des allèles pour ce type d'échantillon en multipliant ces probabilités conditionnelles par la proportion du phénotype dans l'échantillon. Le tableau 4.2 donne cette distribution.

Maintenant, il nous faut revoir les étapes de l'algorithme. Pour ce faire, évaluons d'abord la vraisemblance sur les données complètes pour ce type d'échantillon. Celle-ci est obtenue en conditionnant selon le nombre d'individus cas,

$$\begin{aligned} L_c(V_0, V_1 \mid n_1) &= P(D, \Phi \mid V_0, V_1, n_1) \\ &= \frac{P(D, \Phi, V_0, V_1, n_1)}{P(n_1 \mid V_0, V_1) P(V_0, V_1)} \\ &= \frac{P(D, \Phi, n_1 \mid V_0, V_1)}{P(n_1 \mid V_0, V_1)}. \end{aligned}$$

Puisque les phénotypes déterminent entièrement le nombre de cas dans l'échantillon, on peut retirer n_1 au numérateur. De plus, la probabilité d'obtenir n_1 cas d'un tirage aléatoire simple ne dépend pas des distributions à estimer. Ainsi, en retranchant les

Tableau 4.2 Distribution attendue des allèles au gène causal dans un échantillon à proportion fixée de cas

Nombre de copies du gène causal	Cas	Témoin	total
T=00	$f_0(1-p)^2\frac{\omega}{f}$	$(1-f_0)(1-p)^2\frac{(1-\omega)}{(1-f)}$	q_{00}
T=01	$f_1 p(1-p)\frac{\omega}{f}$	$(1-f_1)p(1-p)\frac{(1-\omega)}{(1-f)}$	q_{01}
T=10	$f_1 p(1-p)\frac{\omega}{f}$	$(1-f_1)p(1-p)\frac{(1-\omega)}{(1-f)}$	q_{10}
T=11	$f_2 p^2\frac{\omega}{f}$	$(1-f_2)p^2\frac{(1-\omega)}{(1-f)}$	q_{11}
total	ω	$(1-\omega)$	1

termes qui ne dépendent pas de V_0 et V_1, on retrouve la vraisemblance sur les données complètes,

$$\begin{aligned} L_c(V_0, V_1 \mid n_1) &= \frac{P(D, \Phi, n_1 \mid V_0, V_1)}{P(n_1)} \\ &= \frac{P(D, \Phi \mid V_0, V_1)}{\binom{n}{n_1} f^{n_1}(1-f)^{(n-n_1)}} \\ &= K(F, p, n_1) \prod_h V_0(h)^{m_{h^0}} V_1(h)^{m_{h^1}}. \end{aligned}$$

Ceci nous permet de conclure que l'étape M de l'algorithme, décrite par les équations (4.2) et (4.3) de la page 65, n'est pas affectée par un échantillonnage à proportion fixée de cas.

Rappelons que l'étape E dépendait des probabilités des diplotypes, conditionnelles au génotype et au phénotype. Démontrons maintenant que ces probabilités ne sont pas modifiées par le type d'échantillonnage. Calculons dans un premier temps la probabilité conjointe d'un diplotype parental et d'un phénotype, conditionnellement à la fréquence ω d'individus cas. On obtient celle-ci en ajoutant une condition sur la proportion d'individus cas à l'équation (4.4) de la page 66.

$$\begin{aligned} P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0, V_1, \omega) &= P(\phi \mid n_1) P(T = \delta_1\delta_2 \mid \phi, \omega) \\ &\times \; P([h_i, h_j] \mid T = \delta_1\delta_2, \omega). \end{aligned}$$

Une fois déterminé le statut T au gène causal, la probabilité du diplotype parental ne dépend que des distributions V_0 et V_1. De même, une fois le phénotype connu, T ne dépend pas de ω. Seule la probabilité de présenter le phénotype est affectée, ce qui nous donne

$$P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0, V_1, \omega) = P(\phi \mid \omega) P(T = \delta_1\delta_2 \mid \phi) V_{\delta_1}(h_i) V_{\delta_2}(h_j).$$

De manière analogue à ce qui a été fait pour un échantillon aléatoire simple, le terme $P(\phi \mid \omega)$ s'annule lors du calcul de la probabilité conditionnelle. On retrouve donc le même résultat que précédemment. Puisque ces probabilités ne sont pas affectées par l'échantillonnage, l'étape E de l'algorithme, décrite à l'équation (4.15) de la page 68,

demeure identique. Finalement, on peut en conclure que l'algorithme EM que nous venons de décrire est directement applicable à un échantillon à proportion fixée de cas, ce qui n'est pas le cas pour les méthodes présentées au chapitre précédent.

4.1.4 Algorithme

Nous venons de développer et décrire un algorithme EM permettant d'estimer les distributions V_0 et V_1. Afin de mieux mettre en évidence la chronologie des étapes, cette section résume ce dernier en le présentant sous une forme plus algorithmique. Les étapes successives sont illustrées à la figure 4.1.

Algorithme EM sur les distributions V_0 et V_1

1. Paramètres
   - Déterminer le modèle de pénétrance $F = (f_0, f_1, f_2)$ à appliquer, ainsi que la fréquence de la mutation causale p. Nous verrons dans la section suivante comment ces paramètres peuvent être estimés, si nécessaire.

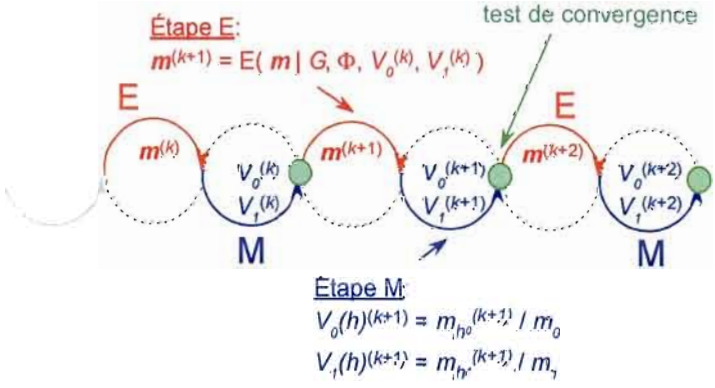

Figure 4.1 Étapes de l'algorithme EM conditionnel aux phénotypes

- Calculer les probabilités $P(T \mid \phi)$ (voir page 66),

$$\begin{aligned}
P(T = 00 \mid \phi = 1) &= \frac{f_0(1-p)^2}{f}, \\
P(T = 01 \mid \phi = 1) &= \frac{f_1 p(1-p)}{f}, \\
P(T = 10 \mid \phi = 1) &= \frac{f_1 p(1-p)}{f}, \\
P(T = 11 \mid \phi = 1) &= \frac{f_2 p^2}{f}, \\
P(T = 00 \mid \phi = 0) &= \frac{(1-f_0)(1-p)^2}{1-f}, \\
P(T = 01 \mid \phi = 0) &= \frac{(1-f_1)p(1-p)}{1-f}, \\
P(T = 10 \mid \phi = 0) &= \frac{(1-f_1)p(1-p)}{1-f}, \\
P(T = 11 \mid \phi = 0) &= \frac{(1-f_2)p^2}{1-f}.
\end{aligned}$$

2. Initialisation

   Choisir des valeurs initiales $V_0^{(0)}$ et $V_1^{(0)}$. Le choix de ces distributions initiales est important, puisqu'il peut influencer le résultat final. Puisque la distribution uniforme représente en quelque sorte la neutralité, cette distribution est suggérée. Les distributions initiales peuvent aussi être choisies de manière à refléter une connaissance préalable de la structure de la population.

3. Étape E

   - Pour chaque génotype g et phénotype ϕ de l'échantillon :

     (a) Calculer, pour tout $[h_i^{\delta_1}, h_j^{\delta_2}] \in g$ (voir page 66),

$$P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0^{(k)}, V_1^{(k)}) = V_{\delta_1}^{(k)}(h_i) V_{\delta_2}^{(k)}(h_j) P(T = \delta_1\delta_2 \mid \phi) P(\phi),$$

     où $P(\phi)$ n'a pas à être évalué explicitement puisqu'il s'annulera lors du calcul de la probabilité conditionnelle.

     (b) Sommer ces probabilités pour obtenir

$$P(g, \phi \mid V_0^{(k)}, V_1^{(k)}) = \sum_{[h_i^{\delta_1}, h_j^{\delta_2}] \in g} P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0^{(k)}, V_1^{(k)}).$$

(c) Effectuer le quotient pour retrouver la probabilité conditionnelle

$$P([h_i^{\delta_1}, h_j^{\delta_2}] \mid g, \phi, V_0^{(k)}, V_1^{(k)}) = \frac{P([h_i^{\delta_1}, h_j^{\delta_2}], \phi \mid V_0^{(k)}, V_1^{(k)})}{P(g, \phi \mid V_0^{(k)}, V_1^{(k)})}.$$

- Calculer, pour toute séquence h^δ (voir page 68),

$$m_{h^\delta}^{(k+1)} = \sum_{(g,\phi)\in(G,\Phi)} 2n_{g,\phi} P([h^\delta, *] \mid g, \phi, V_0^{(k)}, V_1^{(k)}).$$

4. Étape M

   Mettre à jour les distributions en calculant, pour tout h (voir page 65),

$$\begin{aligned} V_0^{(k+1)}(h) &= \frac{m_{h^0}^{(k+1)}}{m_0^{(k+1)}}, \\ V_0^{(k+1)}(h) &= \frac{m_{h^1}^{(k+1)}}{m_1^{(k+1)}}. \end{aligned}$$

5. Test de convergence

   Différents critères peuvent être utilisés. Nous considérerons que la convergence sera atteinte lorsque la plus grande différence absolue $\max_{h,\delta} \mid V_\delta(h)^{(k+1)} - V_\delta^{(k)}(h) \mid$ sera inférieure à une valeur ϵ déterminée par le nombre d'haplotypes possibles et la précision demandée.

   Si $\max_{h,\delta} \mid V_\delta(h)^{(k+1)} - V_\delta^{(k)}(h) \mid < \epsilon$, terminer l'algorithme et reporter comme estimés $\hat{V}_0 = V_0^{(k+1)}$ et $\hat{V}_1 = V_1^{(k+1)}$. Autrement, reprendre à l'étape 3.

4.1.5 Illustration

Considérons une région chromosomique génotypée sur deux marqueurs. Notons par a et A les allèles au premier marqueur et par b et B ceux au deuxième. Nous travaillerons avec un modèle de pénétrance dominant et sans phénocopies, c'est-à-dire que $f_0 = 0$, $f_1 = 1$ et $f_2 = 1$. Soit $p = 0,2$, la fréquence de la mutation causale. De cette population, supposons un échantillon à proportion fixée de cas, de telle sorte que la distribution des génotypes est donnée par le tableau 4.3.

Tableau 4.3 Décompte des génotypes dans un échantillon

Génotype	Phénotype cas	Phénotype témoin
$\{(a,a)(b,b)\}$	0	5
$\{(a,a)(b,B)\}$	6	2
$\{(a,a)(B,B)\}$	2	1
$\{(a,A),(b,b)\}$	0	5
$\{(a,A),(b,B)\}$	5	5
$\{(a,A),(B,B)\}$	2	0
$\{(A,A),(b,b)\}$	0	2
$\{(A,A),(b,B)\}$	3	0
$\{(A,A),(B,B)\}$	2	0
total	20	20

1. Paramètres

   Puisque les paramètres sont connus, calculons dans un premier temps les probabilités $P(T \mid \phi)$ en appliquant les équations (4.5) à (4.12), page 66 :

$$
\begin{aligned}
P(T = 00 \mid \phi = 1) &= 0,0; \\
P(T = 01 \mid \phi = 1) &= 0,\bar{4}; \\
P(T = 10 \mid \phi = 1) &= 0,\bar{4}; \\
P(T = 11 \mid \phi = 1) &= 0,\bar{1}; \\
P(T = 00 \mid \phi = 0) &= 1,0; \\
P(T = 01 \mid \phi = 0) &= 0,0; \\
P(T = 10 \mid \phi = 0) &= 0,0; \\
P(T = 11 \mid \phi = 0) &= 0,0.
\end{aligned}
$$

2. Initialisation

   Il nous faut maintenant attribuer des valeurs initiales aux distributions. Puisque nous n'avons aucune information préalable sur les fréquences attendues, nous prendrons des distributions uniformes. Le tableau 4.4 présente les distributions initiales.

3. Étape E

   Pour l'étape E, il nous faut évaluer les distributions conditionnelles aux génotypes et phénotypes. Considérons le génotype $\{(a,a)(b,B)\}$. Les seuls diplotypes parentaux compatibles sont $[ab, aB]$ et $[aB, ab]$.

   Pour un individu cas, on a alors

$$\begin{aligned}
P([ab^0, aB^1], \phi = 1 \mid V_0^{(0)}, V_1^{(0)}) &= V_0^{(0)}(ab) V_1^{(0)}(aB) P(T = 01 \mid \phi = 1) P(\phi = 1) \\
&= 0,25 \times 0,25 \times 0,444 P(\phi = 1) \\
&= 0,2775\ P(\phi = 1); \\
P([ab^1, aB^0], \phi = 1 \mid V_0^{(0)}, V_1^{(0)}) &= V_1^{(0)}(ab) V_0^{(0)}(aB) P(T = 10 \mid \phi = 1) P(\phi = 1) \\
&= 0,25 \times 0,25 \times 0,444\ P(\phi = 1) \\
&= 0,2775\ P(\phi = 1); \\
P([ab^1, aB^1], \phi = 1 \mid V_0^{(0)}, V_1^{(0)}) &= V_1^{(0)}(ab) V_1^{(0)}(aB) P(T = 11 \mid \phi = 1) P(\phi = 1) \\
&= 0,25 \times 0,25 \times 0,111\ P(\phi = 1) \\
&= 0,006\ 94\ P(\phi = 1);
\end{aligned}$$

Tableau 4.4 Distributions initiales

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

de manière symétrique,

$$\begin{aligned}
P([aB^1, ab^0], \phi = 1 \mid V_0^{(0)}, V_1^{(0)}) &= V_1^{(0)}(aB)V_0^{(0)}(ab)P(T = 10 \mid \phi = 1)P(\phi = 1) \\
&= 0,25 \times 0,25 \times 0,444\ P(\phi = 1) \\
&= 0,2775\ P(\phi = 1); \\
P([aB^0, ab^1], \phi = 1 \mid V_0^{(0)}, V_1^{(0)}) &= V_0^{(0)}(aB)V_1^{(0)}(ab)P(T = 01 \mid \phi = 1)P(\phi = 1) \\
&= 0,25 \times 0,25 \times 0,444\ P(\phi = 1) \\
&= 0,2775\ P(\phi = 1); \\
P([aB^1, ab^1], \phi = 1 \mid V_0^{(0)}, V_1^{(0)}) &= V_1^{(0)}(aB)V_1^{(0)}(ab)P(T = 11 \mid \phi = 1)P(\phi = 1) \\
&= 0,25 \times 0,25 \times 0,111\ P(\phi = 1) \\
&= 0,006\ 94\ P(\phi = 1).
\end{aligned}$$

En effectuant la somme, on obtient pour ce profil,

$$P(\{(a,a)(b,B)\}, \phi = 1 \mid V_0^{(0)}, V_1^{(0)})P(\phi = 1) = 1,123\ 88\ P(\phi = 1).$$

Ce qui nous donne les probabilités conditionnelles :

$$\begin{aligned}
P([ab^0, aB^1] \mid \{(a,a)(b,B)\}, \phi = 1, V_0^{(0)}, V_1^{(0)}) &= 0,246\ 91; \\
P([ab^1, aB^0] \mid \{(a,a)(b,B)\}, \phi = 1, V_0^{(0)}, V_1^{(0)}) &= 0,246\ 91; \\
P([ab^1, aB^1] \mid \{(a,a)(b,B)\}, \phi = 1, V_0^{(0)}, V_1^{(0)}) &= 0,006\ 18; \\
P([aB^1, ab^0] \mid \{(a,a)(b,B)\}, \phi = 1, V_0^{(0)}, V_1^{(0)}) &= 0,246\ 91; \\
P([aB^0, ab^1] \mid \{(a,a)(b,B)\}, \phi = 1, V_0^{(0)}, V_1^{(0)}) &= 0,246\ 91; \\
P([aB^1, ab^1] \mid \{(a,a)(b,B)\}, \phi = 1, V_0^{(0)}, V_1^{(0)}) &= 0,006\ 18.
\end{aligned}$$

Nous avons illustré ici le calcul pour un seul profil. L'application de l'algorithme nécessite de répéter l'opération pour tous les génotypes et phénotypes.

Les probabilités précédentes nous permettent de calculer que, pour six individus cas présentant ce génotype, il y en aura en moyenne $6 \times (0,246\ 91 + 0,006\ 18) = 1,518\ 54$ qui porteront la séquence ab^1 sur l'haplotype maternel et le même nombre sur l'haplotype paternel. Le nombre moyen de séquences de type ab^1 pour ces

individus est donc de 3,037 08. Le tableau 4.5 donne les fréquences moyennes des haplotypes obtenues en répétant l'opération pour tous les génotypes et phénotypes.

4. Étape M

   Une fois les fréquences moyennes obtenues, l'étape M consiste à actualiser les distributions. Par exemple, on retrouve facilement

$$\begin{aligned} V_0(ab)^{(1)} &= \frac{m_{ab^0}^{(1)}}{m_0^{(1)}} \\ &= \frac{11,638\ 90}{28,888\ 90} \\ &= 0,402\ 88. \end{aligned}$$

   On déduit les autres paramètres de manière analogue, ce qui nous permet d'obtenir les distributions données au tableau 4.6.

5. Test de convergence

   Supposons que l'on souhaite atteindre un critère de convergence de $\epsilon \leq 0,000\ 01$. Pour chaque séquence h, on évalue les différences $\mid V_0^{(1)}(h) - V_0^{(0)}(h) \mid$ et $\mid V_1^{(1)}(h) - V_1^{(0)}(h) \mid$. Ici, la différence maximale est donnée par $\mid V_0^{(1)}(ab) - V_0^{(0)}(ab) \mid = 0,150288$. Cette différence étant supérieure à $0,000\ 01$, on rejette le test de convergence. On reprend alors l'algorithme à l'étape E.

L'algorithme converge finalement vers les distributions données au tableau 4.7.

Tableau 4.5 Fréquences moyennes des haplotypes à la première itération

Haplotype h	Fréquence $m_{h^0}^{(1)}$	Fréquence $m_{h^1}^{(1)}$
ab	11,638 90	2,361 11
aB	6,472 22	4,027 78
Ab	6,972 22	1,527 78
AB	3,805 56	3,194 44
total	28,888 90	11,111 11

Tableau 4.6 Distributions après la première itération

Haplotype h	Distribution $V_0^{(1)}(h)$	Distribution $V_1^{(1)}(h)$
ab	0,402 88	0,212 50
aB	0,224 04	0,362 50
Ab	0,241 35	0,137 50
AB	0,131 73	0,287 50

Tableau 4.7 Distributions estimées

Haplotype h	Distribution $\hat{V}_0(h)$	Distribution $\hat{V}_1(h)$
ab	0,480 99	0,000 77
aB	0,166 53	0,521 35
Ab	0,299 12	0,000 17
AB	0,053 36	0,477 71

4.2 Évaluation des paramètres

4.2.1 Fréquence du gène causal

L'algorithme que nous avons construit suppose la connaissance du modèle de pénétrance et de la proportion d'haplotypes mutants dans la population. En pratique, même en supposant le modèle de pénétrance connu, on ignore souvent la valeur de la fréquence du gène causal p, mais on connaît assurément la fréquence f du caractère dans la population. On constate que f dépend entièrement du modèle de pénétrance et de la fréquence du gène causal. Réciproquement, on peut aussi obtenir p à partir de f et du modèle de pénétrance. On a

$$f = f_0(1-p)^2 + 2f_1p(1-p) + f_2p^2$$

d'où l'équation quadratique en p suivante :

$$0 = (f_0 + f_2 - 2f_1)p^2 + 2(f_1 - f_0)p + (f_0 - f)$$

que l'on peut aisément résoudre. Ce problème se décompose en deux cas. Premièrement, si $(f_0 + f_2 - 2f_1) = 0$, l'équation précédente se réduit à un premier degré et on a directement :

$$p = \frac{f - f_0}{2(f_1 - f_0)} \quad \text{si } f_0 + f_2 - 2f_1 = 0.$$

Autrement, on applique la formule des zéros d'un polynôme du second degré :

$$\Delta = 4[(f_1 - f_0)^2 - (f_0 - f)(f_0 - 2f_1 + f_2)],$$

$$\begin{aligned} p &= \frac{2(f_0 - f_1) \pm \sqrt{\Delta}}{2(f_0 + f_2 - 2f_1)} \\ &= \frac{f_0 - f_1 \pm \sqrt{f_1^2 - f_0 f_1 + f(f_0 - 2f_1 + f_2)}}{(f_0 + -2f_1 + f_2)}, \end{aligned} \tag{4.16}$$

pour laquelle il y a toujours au moins une solution comprise entre 0 et 1 si f est compatible avec le modèle de pénétrance. Supposons que $0 \leq f_0 \leq f_1 \leq f_2 \leq 1$. Notons que cette hypothèse plausible repose sur le principe que la probabilité d'observer un caractère d'origine génétique augmente avec le nombre d'allèles en cause portées par l'individu. Dans ce cas, il existe une unique solution $0 < p < 1$ à l'équation (4.16). Démontrons ceci par l'absurde. Supposons qu'il existe deux solutions distinctes $0 < p_0 < 1$ et $0 < p_1 < 1$ correspondant respectivement à la racine positive et négative de Δ. Dans ce cas on obtient que $0 < p_0 + p_1 < 2$ et alors,

$$p_0 + p_1 = 2\frac{f_0 - f_1}{f_0 - 2f_1 + f_2},$$

d'où

$$0 < \frac{f_0 - f_1}{f_0 - 2f_1 + f_2} < 1.$$

Puisque, par hypothèse, $f_0 \leq f_1$ et que le quotient est positif, on a nécessairement que $f_0 - 2f_1 + f_2 < 0$. Dans ce cas,

$$\begin{aligned} f_0 - f_1 &> f_0 - 2f_1 + f_2, \\ 0 &> f_2 - f_1, \end{aligned}$$

d'où $f_2 < f_1$, ce qui contredit les hypothèses. Il n'existe donc pas deux solutions distinctes admissibles comme valeurs de p lorsque $0 \leq f_0 \leq f_1 \leq f_2 \leq 1$. Pour d'autres

types de modèles, si on suppose que le gène causal est rare, on prendra la plus petite des solutions admissibles.

En supposant le modèle de pénétrance connu, il nous est donc possible de calculer de manière exacte les probabilités $P(T \mid \phi)$. En pratique toutefois, il n'est pas toujours évident de déterminer F. Il faudra alors l'estimer.

4.2.2 Estimation du modèle de pénétrance

Jusqu'à présent, nous avons supposé le modèle de pénétrance connu. Toutefois, c'est rarement le cas, ce qui rend nécessaire son estimation. En théorie, ce modèle, ainsi que la probabilité p, peuvent être estimés en même temps que les distributions d'haplotypes par l'algorithme EM. Cette approche a d'ailleurs été exploitée par Ito *et al.* (2004). Leur méthodologie diffère largement de celle présentée ici en ce sens qu'ils estiment les paramètres d'une unique distribution V pour les haplotypes, desquels certains sont présupposés en lien avec le caractère, donc porteurs du gène. Les résultats publiés portent sur un modèle particulier dominant et 6 marqueurs de type SNP. Ils ont obtenu de bons résultats lorsque le modèle était supposé de type dominant, donc à deux paramètres, que le groupe des haplotypes considérés porteurs était le bon et que ceux-ci n'étaient pas trop rares. Dans le cas qui nous concerne, on ne fait aucune supposition quant à l'identité des séquences haploïdes mutantes ou le type de modèle. On rencontre donc des obstacles qui rendent l'estimation simultanée plus hasardeuse, quoique possible.

Un premier obstacle se pose quant aux contraintes à appliquer au modèle. Tous les modèles ne sont pas compatibles avec la fréquence du caractère dans la population. Aussi, si l'on dispose d'une certaine connaissance de la maladie, il est possible que l'on ait une idée approximative du modèle de pénétrance, ou de la fréquence de la mutation. Enfin, on accepte généralement que la probabilité de présenter le caractère est plus grande pour un individu portant le gène causal, c'est-à-dire $f_0 \leq f_1 \leq f_2$. Or, il est ardu d'intégrer ce type de contraintes lors de l'optimisation, d'autant plus si elles

varient d'une expérimentation à l'autre. Notons aussi que l'estimation des paramètres du modèle ne serait pas robuste à un échantillonnage à proportion fixée de cas.

Le second obstacle qui se présente est lié à l'algorithme EM lui-même. Rappelons que l'algorithme EM converge vers un maximum local de vraisemblance qui dépend du point de départ. Lorsqu'on suppose le modèle inconnu, on augmente le risque de converger vers un maximum local très éloigné du maximum global. Ceci est d'autant plus vrai si l'espace des paramètres est très grand. Ainsi, pour L marqueurs génotypés, il existe 2^L haplotypes possibles. Par exemple, avec seulement 10 marqueurs, on se retrouve avec plus de 1000 paramètres à estimer par distribution (contre 64 avec 6 marqueurs). Si, en plus, le déséquilibre de liaison n'est pas particulièrement fort, le modèle aura un plus grand impact sur la vraisemblance que les distributions V_0 et V_1. Sans entrer dans de laborieux calculs d'analyse, on peut penser que les paramètres du modèle convergeront plus rapidement que les distributions. Par conséquent, l'estimation obtenue dépendra grandement des valeurs initiales attribuées à V_0 et V_1. Une façon de contourner ce problème serait de considérer un très grand nombre de valeurs initiales. Considérant l'espace sur les paramètres, cette solution est difficilement applicable.

Étant donné les risques liés à l'estimation simultanée des paramètres, nous avons développé une méthode permettant d'appliquer un certain contrôle sur l'estimation du modèle. Étudions d'abord le cas d'un échantillon aléatoire simple. Considérons un modèle de pénétrance F et une valeur de p compatibles avec la population, c'est-à-dire que $f = f_0(1-p)^2 + 2f_1p(1-p) + f_2p^2$ où f est une constante connue. La vraisemblance de F, p, V_0 et V_1 sur l'échantillon ordonné est

$$\begin{aligned} L(F,p,V_0,V_1) &= P(G,\Phi \mid F,p,V_0,V_1) \\ &= \prod_{(g,\phi)} P(g,\phi \mid F,p,V_0,V_1)^{n_{(g,\phi)}}. \end{aligned}$$

Si l'échantillon n'est pas aléatoire simple, mais plutôt à proportion fixée d'individus cas, il suffit de calculer la vraisemblance conditionnelle au nombre fixé de cas :

$$\begin{aligned} L(F,p,V_0,V_1 \mid n_1) &= P(G,\Phi \mid F,p,V_0,V_1,n_1) \\ &= \frac{P(G,\Phi,n_1 \mid F,p,V_0,V_1)}{P(n_1 \mid F,p,V_0,V_1)} \\ &= \frac{P(G,\Phi \mid F,p,V_0,V_1)}{P(n_1 \mid F,p,V_0,V_1)} \\ &= \frac{\prod_{(g,\phi)} P(g,\phi \mid F,p,V_0,V_1)^{n_{(g,\phi)}}}{\binom{n}{n_1} f^{n_1}(1-f)^{n-n_1}}. \end{aligned}$$

Puisque n, f, n_1 et les $n_{g,\phi}$ sont des constantes pour l'échantillon, la vraisemblance peut alors être exprimée à une constante proportionnelle près

$$L(F,p,V_0,V_1) \propto \prod_{(g,\phi)} P(g,\phi \mid F,p,V_0,V_1)^{n_{(g,\phi)}}.$$

Ainsi, si les distributions V_0 et V_1 étaient connues, il serait possible d'estimer le modèle de pénétrance par la méthode du maximum de vraisemblance en optimisant l'expression précédente sur un ensemble de modèles $\mathcal{F}$.

En pratique, les distributions V_0 et V_1 ne sont pas connues. Pour cette raison, nous chercherons plutôt à maximiser la vraisemblance conjointe de F, p, V_0 et V_1. Cette vraisemblance atteindra son maximum au point $(\hat{F},\hat{p},\hat{V}_0,\hat{V}_1)$ tel que les distributions $\hat{V}_0$ et $\hat{V}_1$ sont les estimateurs à maximum de vraisemblance pour V_0 et V_1, conditionnellement à $\hat{F}$ et $\hat{p}$. Il est alors possible d'obtenir les estimateurs $\hat{F}$ et $\hat{p}$ en maximisant

$$\hat{L}(F,p) = \prod_{(g,\phi)} P(g,\phi \mid F,p,\hat{V}_{0\,|(F,p)},\hat{V}_{1\,|(F,p)})^{n_{(g,\phi)}}, \tag{4.17}$$

où l'on obtiendra $\hat{V}_{0\,|(F,p)}$ et $\hat{V}_{1\,|(F,p)}$ par l'algorithme EM conditionnel aux phénotypes. Selon les contraintes que l'on souhaite appliquer au modèle, on choisira un espace plus ou moins large pour les valeurs admissibles de F et p.

4.2.3 Algorithme d'estimation du modèle

Dans la section précédente, nous avons proposé un algorithme permettant d'estimer le modèle de pénétrance. Nous en résumons ici les principales étapes.

Algorithme d'estimation du modèle

1. Choisir un ensemble fini de modèles de pénétrance à considérer $\mathcal{F}$, ainsi qu'un intervalle $I \subseteq]0,1[$ pour la fréquence de la mutation.

   Cet intervalle représente une information préalable sur la fréquence de la mutation causale. Ainsi, si on suppose que celle-ci est plutôt rare, on pourrait choisir un intervalle situé entre 0,01 et 0,1.

2. Pour chaque modèle $F \in \mathcal{F}$:

   (a) Calculer la valeur du paramètre p (page 79) :

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

   (b) Si $p \in I$, obtenir les estimateurs $\hat{V}_{0|F,p}$ et $\hat{V}_{1|F,p}$ par l'algorithme EM conditionnel au phénotype. Sinon, éliminer ce modèle et retourner à l'étape 2.

   (c) Estimer la vraisemblance du modèle en calculant $\hat{L}(F,p)$ (voir equation (4.17)),

   $$\hat{L}(F,p) = \prod_{(g,\phi)} P(g,\phi \mid F, p, \hat{V}_{0\ |(F,p)}, \hat{V}_{1\ |(F,p)})^{n_{(g,\phi)}}.$$

3. Choisir comme estimateur le modèle tel que $\hat{L}(F,p)$ est maximal.

Nous avons développé dans ce chapitre un algorithme EM conditionnel aux phénotypes permettant d'estimer les séquences haploïdes associées à un échantillon diploïde. Cet algorithme nous permet aussi d'estimer le statut au gène causal, ce qui est nécessaire pour l'application de la méthode de cartographie que nous voulons généraliser. Cette généralisation sera d'ailleurs traitée au prochain chapitre.

CHAPITRE V

GÉNÉRALISATION DE MAPARG

La version originale de la méthode de cartographie présentée dans Larribe *et al.* (2002) est basée sur un modèle d'échantillonnage pondéré nécessitant un échantillon de séquences haploïdes. Dans le cas d'individus diploïdes, on connaît rarement les séquences formant le diplotype. De plus, il est à peu près impossible de déterminer avec certitude le nombre d'allèles mutants portés par un individu sur le gène causal, et encore moins leur répartition sur les deux haplotypes. De ce fait, la méthode n'est pas applicable directement pour des individus diploïdes. Nous présenterons ici deux façons de traiter ce problème. La première consiste à ajouter des étapes au modèle d'échantillonnage pondéré, de manière à estimer la vraisemblance sur les génotypes et phénotypes des individus. La seconde consiste à réduire l'échantillon diploïde à des paramètres haploïdes. Notons qu'un résumé des notations utilisées dans ce chapitre est présenté à l'annexe D.

5.1 Vraisemblance sur les génotypes et phénotypes

5.1.1 Modèle d'échantillonnage pondéré

Pour des raisons de calculs, considérons l'ensemble des individus comme étant ordonné. Ainsi, on calculera $Q_{r_T}(G, \Phi)$ comme étant la probabilité d'obtenir les génotypes et phénotypes des individus dans un ordre précis. Cette façon de faire simplifie les calculs en nous évitant de considérer des constantes multinomiales. Théoriquement, il serait possible d'évaluer $Q_{r_T}(G, \Phi)$ en sommant sur tous les ensembles H_0 compatibles. Ainsi,

on aurait

$$Q_{r_T}(G,\Phi) = \sum_{H_0} Q_{r_T}(G,\Phi \mid H_0) Q_{r_T}(H_0).$$

La difficulté repose ici dans l'évaluation de $Q_{r_T}(G,\Phi \mid H_0)$, qui nécessite l'énumération de toutes les façons d'agencer les séquences de H_0 en diplotypes pour former les génotypes observés. Une façon de le faire consiste à insérer une étape supplémentaire entre les génotypes et les séquences haploïdes.

Soit H_{-1}, l'ensemble ordonné des diplotypes parentaux des individus. Remarquons que la différence entre H_{-1} et H_0 réside dans le fait que le premier échantillon associe un doublet de séquences haploïdes à chaque individu, ce que ne fait pas le second. Notons aussi par H_{-2}, les génotypes et phénotypes. La figure 5.1 illustre la différence entre ces nouvelles étapes. Nous pouvons alors augmenter l'équation de récurrence (2.1) à la page 19, de sorte que, pour tout $\tau \geq -2$,

$$Q_{r_T}(H_\tau) = \sum_{H_{\tau+1}} Q_{r_T}(H_\tau \mid H_{\tau+1}) Q_{r_T}(H_{\tau+1}).$$

En reprenant les étapes et notations données au second chapitre, on obtient alors le modèle d'échantillonnage pondéré :

$$\begin{aligned}
Q_{r_T}(H_{-2}) &= \sum_{H_{-1}} \sum_{H_0}, \ldots, \sum_{H_{\tau^*-1}} \prod_{\tau=-2}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) P_{r_{T_0}}(H_{\tau+1} \mid H_\tau) \\
&= \sum_{H_{-1}} \sum_{H_0}, \ldots, \sum_{H_{\tau^*-1}} \left[\prod_{\tau=-2}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) \right] \left[\prod_{\tau=-2}^{\tau^*-1} P_{r_{T_0}}(H_{\tau+1} \mid H_\tau) \right] \\
&= E_{P_{r_{T_0}}} \left[\prod_{\tau=-2}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) \right],
\end{aligned}$$

où

$$h_{r_T r_{T_0}}(H_\tau, H_{\tau+1}) = \frac{Q_{r_T}(H_\tau \mid H_{\tau+1})}{P_{r_{T_0}}(H_{\tau+1} \mid H_\tau)}.$$

Ainsi, l'intégration de la réalité diploïde se résume alors à ajouter deux étapes au modèle d'échantillonnage pondéré présenté précédemment. De ce fait, la simulation de diplotypes et de graphes selon une distribution $P_{r_{T_0}}$ permet d'estimer la vraisemblance sur les génotypes, $L(r_T) = Q_{r_T}(H_{-2})$, par une moyenne des valeurs de $\prod_{\tau=-2}^{\tau^*-1} h_{r_T r_{T_0}}(H_\tau, H_{\tau+1})$.

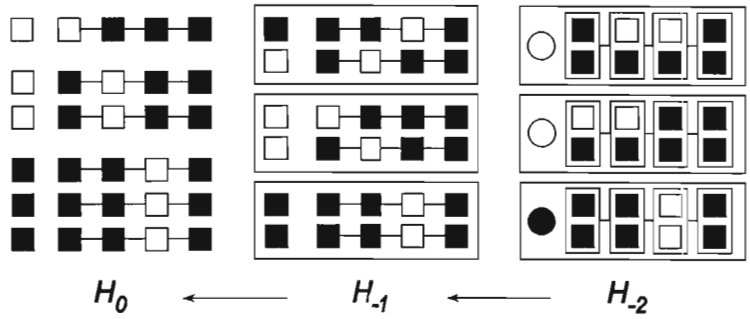

Figure 5.1 Étapes H_0, H_{-1} et H_{-2}. À droite, on a les génotypes et phénotypes de trois individus : les carrés représentent les allèles aux marqueurs, tandis que les cercles représentent les phénotypes (noir pour cas, blanc pour témoin). Au centre, les haplotypes de chaque individu sont connus et sont représentés par deux chaînes de marqueurs. Le carré le plus à gauche représente la mutation causale (noir pour mutant, blanc pour normal). À gauche, on retrouve la liste des séquences H_0.

Les distributions Q et P ont été données au second chapitre pour $\tau \geq 0$. Il nous reste donc à développer les distributions manquantes.

5.1.2 Distributions

Le modèle d'échantillonnage pondéré que nous venons de décrire nécessite la connaissance des distributions $Q_{r_T}(H_{-2} \mid H_{-1})$, $Q_{r_T}(H_{-1} \mid H_0)$, $P_{r_{T_0}}(H_{-1} \mid H_{-2})$ et $P_{r_{T_0}}(H_0 \mid H_{-1})$, les autres étant connues.

Notons dans un premier temps que ces distributions ne dépendent pas de la position de la mutation causale. Pour cette raison, nous omettrons l'indice r_T dans les notations. Pour débuter, notons que

$$P(H_0 \mid H_{-1}) = 1. \tag{5.1}$$

En effet, les diplotypes formant H_{-1} déterminent de manière unique l'ensemble des séquences haploïdes. Étant donné que l'on a supposé que les génotypes étaient ordonnés, la distribution $P(H_{-1} \mid H_{-2})$ est entièrement déterminée par l'algorithme servant à la reconstruction. Nous avons choisi ici d'utiliser un algorithme EM conditionnel aux phénotypes, mais un autre algorithme aurait aussi pu être utilisé. Si l'on construit les

diplotypes des individus de manière indépendante, $P(H_{-1} \mid H_{-2})$ est donnée par le produit des probabilités marginales des individus. Ainsi,

$$P(H_{-1} \mid H_{-2}) = \prod_i P(d_i \mid g_i, \phi_i, \hat{V}_0, \hat{V}_1), \tag{5.2}$$

où d_i représente le diplotype parental $[h_1^{\delta_1}, h_2^{\delta_2}]$ associé à l'individu i dans la reconstruction H_{-1}. Rappelons que les probabilités $P([h_1^{\delta_1}, h_2^{\delta_2}] \mid g_i, \phi_i, \hat{V}_0, \hat{V}_1)$ sont données au chapitre précédent par le quotient des équations (4.4) et (4.13), page 66.

Les distributions $Q(H_{-2} \mid H_{-1})$ et $Q(H_{-1} \mid H_{-0})$ dépendent quant à elles du type d'échantillon considéré, soit aléatoire simple ou avec une proportion fixée d'individus cas.

5.1.3 Échantillon aléatoire simple

Considérons un échantillon aléatoire simple. Les génotypes étant entièrement déterminés par les diplotypes parentaux, la probabilité d'observer les génotypes et phénotypes des individus ne dépend que du modèle de pénétrance F et des allèles au gène causal. Ainsi,

$$Q(H_{-2} \mid H_{-1}) = \prod_i P(\phi_i \mid T_i), \tag{5.3}$$

où T_i est le nombre de copies de la mutation causale portées par l'individu i dans la reconstruction H_{-1}. Rappelons que les probabilités $P(\phi \mid T)$ sont données directement par le modèle de pénétrance. Ainsi, si l'individu i est cas, $\phi_i = 1$ et $P(\phi_i \mid T_i) = f_{T_i}$. S'il est témoin, on aura plutôt $\phi_i = 0$ et $P(\phi_i \mid T_i) = (1 - f_{T_i})$. On peut donc réécrire l'équation précédente sous la forme :

$$Q(H_{-2} \mid H_{-1}) = \prod_i f_{T_i}^{(\phi_i)} (1 - f_{T_i})^{(1-\phi_i)}.$$

Considérons maintenant la distribution $Q(H_{-1} \mid H_0)$. Puisque l'échantillon est aléatoire simple, $Q(H_{-1} \mid H_0)$ est donnée par la probabilité de jumeler aléatoirement les éléments de H_0 pour obtenir H_{-1}, de manière à associer les bonnes paires aux bons individus. Le calcul de cette probabilité est plus aisé si on considère les permutations sur les éléments

de H_0. Regardons H_{-1} comme une liste pour laquelle les deux premiers éléments forment une paire ordonnée et ainsi de suite pour les éléments suivants. Puisqu'il existe une unique façon de lister les éléments de H_0 pour obtenir H_{-1}, $Q(H_{-1} \mid H_0)$ sera donnée par l'inverse du nombre de permutations distinctes des éléments de H_0.

Calculons le nombre de manières distinctes d'ordonner les éléments de H_0. Notons par m_s la fréquence de la séquence s dans H_0. Pour chaque ordre distinct, il y a $\prod_s m_s!$ façons de permuter les séquences identiques sans affecter le résultat. Puisqu'il y a $(2n)!$ façons de permuter les $2n$ éléments de H_0, il y a

$$\frac{(2n)!}{\prod_s m_s!}$$

façons distinctes d'ordonner les éléments de H_0.

L'inverse de cette valeur nous donne la probabilité d'obtenir une reconstruction équivalente à H_{-1} en ordonnant les éléments de la liste H_0 de manière aléatoire. On obtient alors que

$$Q(H_{-1} \mid H_0) = \frac{\prod_s m_s!}{(2n)!}. \tag{5.4}$$

L'exemple suivant illustre ce calcul.

Considérons les paires $\{aa, ab, bc, ba\}$ auxquelles correspond l'ensemble d'éléments $\{a, a, a, a, b, b, b, c\}$. Si on effectue le calcul selon un modèle d'urne, on obtient :

$$\begin{aligned} Prob(aa) &= \frac{4}{8}\frac{3}{7}, \\ Prob(ab \mid aa) &= \frac{2}{6}\frac{3}{5}, \\ Prob(bc \mid aa, ab) &= \frac{2}{4}\frac{1}{3}, \\ Prob(ba \mid aa, ab, bc) &= 1/2, \end{aligned}$$

ce qui nous donne une probabilité de $1/280$ de former cet ensemble ordonné de paires.

Si on effectue maintenant le calcul selon la méthode démontrée précédemment, on obtient que $n = 4, m_1 = 4, m_2 = 3, m_3 = 1$. Il y a donc $4! \times 3! \times 1! = 144$ façons de permuter les séquences identiques entre elles, ce qui nous donne $(2 \times 4)!/144 = 280$ manières dis-

tinctes d'ordonner les séquences. La probabilité d'obtenir l'ensemble de paires est bel et bien donnée par 1/280.

Ces deux calculs sont de fait équivalents. Cependant, pour de gros échantillons ou des applications informatiques, le modèle d'urne devient plus lourd à appliquer.

5.1.4 Échantillon à proportion fixée d'individus cas

Considérons maintenant un échantillon d'individus pour lequel la proportion d'individus cas est fixée, c'est-à-dire qu'on a choisis aléatoirement un nombre n_1 d'individus parmi les cas dans la population, et $n - n_1$ individus parmi les témoins. La proportion de cas dans l'échantillon est alors donnée par $\omega = n_1/n$. La distribution attendue des allèles pour ce type d'échantillon est donnée au chapitre précédent par le tableau 4.2, page 69. Notons

$$\begin{aligned}
\alpha_0 &= \frac{f_0\omega/f}{f_0\omega/f + (1-f_0)(1-\omega)/(1-f)}, \\
\alpha_1 &= \frac{f_1\omega/f}{f_1\omega/f + (1-f_1)(1-\omega)/(1-f)}, \\
\alpha_2 &= \frac{f_2\omega/f}{f_2\omega/f + (1-f_2)(1-\omega)/(1-f)},
\end{aligned}$$

où α_i représente la proportion attendue d'individus présentant le caractère parmi ceux portant i copies de la mutation causale dans l'échantillon. Ces proportions ont été déduites du tableau 4.2, page 69, par le ratio des fréquences de cas et du nombre correspondant d'individus respectivement non porteurs, simplement porteurs et doublement porteurs. Soient $q_0 = q_{00}$, $q_1 = q_{01} + q_{10}$ et $q_2 = q_{11}$, respectivement la fréquence d'individus non porteurs, simplement porteurs et doublement porteurs. Enfin, notons par q la fréquence attendue de la mutation causale. Cette valeur est obtenue par une moyenne sur les individus :

$$q = q_1 + 2q_2.$$

Afin de permettre les calculs, nous supposerons que l'échantillon d'individus a été choisi de manière aléatoire simple dans une sous population présentant ces paramètres. On

peut alors calculer $Q(H_{-2} \mid H_{-1})$ sensiblement comme on l'a fait à la section précédente, c'est-à-dire :

$$Q(H_{-2} \mid H_{-1}) = \prod_{i=1}^{n} (\alpha_{T_i})^{\Phi_i} (1 - \alpha_{T_i})^{1-\Phi_i}.$$

Considérons maintenant la distribution $Q(H_{-1} \mid H_0)$. En raison de la structure de la sous-population, on ne peut plus considérer que les séquences se pairent de manière tout à fait aléatoire pour former les individus. Notons par N_0, N_1 et N_2 le nombre total de diplotypes de H_{-1} portant respectivement 0, 1 et 2 copies du gène causal. Calculons en premier lieu la probabilité d'obtenir cette distribution à partir de H_0. Soit $M = 2N_2 + N_1$ le nombre de séquences porteuses dans H_0. On cherche alors $Q(N_0, N_1, N_2 \mid M)$.

Dans un échantillon aléatoire simple provenant de la sous-population, la probabilité d'observer M séquences porteuses parmi n individus diploïdes est simplement donnée par une distribution binomiale :

$$Q(M) = \binom{2n}{M} q^M (1-q)^{2n-M}.$$

De manière analogue, N_0, N_1 et N_2 suivent une distribution multinomiale, ce qui nous donne :

$$Q(N_0, N_1, N_2, M) = \frac{n!}{N_0! N_1! N_2!} q_0^{N_0} q_1^{N_1} q_2^{N_2}.$$

En effectuant le quotient, on se retrouve avec la probabilité conditionnelle :

$$Q(N_0, N_1, N_2 \mid M) = \left[\frac{n!}{N_0! N_1! N_2!}\right] \left[\frac{(2n-M)! M!}{(2n!)}\right] \left[\frac{q_0^{N_0} q_1^{N_1} q_2^{N_2}}{q^M (1-q)^{2n-M}}\right].$$

On peut maintenant calculer $Q(H_{-1} \mid N_0, N_1, N_2, H_0)$. Pour ce faire, nous devons dénombrer les permutations distinctes sur les éléments de H_0 permettant de former les bonnes fréquences d'individus non porteurs, simplement porteurs et doublement porteurs. Premièrement, $2N_0$ séquences non mutantes doivent être choisies parmi les $(2n - M) = (2N_0 + N_1)$ que contient la liste H_0, pour former les individus non porteurs. Il y a

$$\begin{aligned} \binom{2n-M}{2N_0} &= \frac{(2n-M)!}{(2N_0)!(2n-M-2N_0)!} \\ &= \frac{(2n-M)!}{(2N_0)! N_1!} \end{aligned}$$

façons de le faire. Ensuite, il y a

$$\frac{(2N_0)!}{N_0!}$$

manières de former des diplotypes parentaux en jumelant ces $2N_0$ séquences, pour un total de

$$\frac{(2n-M)!}{N_1!N_0!}$$

combinaisons. On peut procéder à un raisonnement similaire pour les N_2 individus doublement porteurs, ce qui nous donne, après simplification,

$$\frac{M!}{N_1!N_2!}$$

possibilités. Il reste alors $2^{N_1}N_1!$ manières de jumeler les autres séquences pour former les individus porteurs d'une unique copie du gène causal. Enfin, il y a $n!$ permutations possibles sur l'ensemble des couples obtenus. En corrigeant pour les séquences identiques, on obtient

$$\frac{(2n-M)!}{N_1!N_0!}\frac{M!}{N_2!N_1!}\frac{n!N_1!2^{N_1}}{\prod_s(m_s!)}$$

manières distinctes d'ordonner les éléments de H_0 pour former N_0, N_1 et N_2 individus portant respectivement 0, 1 et 2 copies de la mutation causale. Or, une seule de ces listes permet d'obtenir la reconstruction H_{-1}. L'inverse de l'équation précédente nous donne alors, après simplification,

$$Q(H_{-1} \mid N_0, N_1, N_2, H_0) = \frac{\prod_s(m_s!)N_0!N_1!N_2!}{n!(2n-M)!M!2^{N_1}}.$$

Finalement, on déduit la probabilité recherchée en combinant les résultats précédents :

$$\begin{aligned} Q(H_{-1} \mid H_0) &= Q(H_{-1} \mid N_0, N_1, N_2, H_0)Q(N_0, N_1, N_2 \mid M) \\ &= \frac{\prod_s(m_s!)N_0!N_1!N_2!}{n!(2n-M)!M!2^{N_1}}\frac{n!}{N_0!N_1!N_2!}\frac{(2n-M)!M!}{(2n!)}\frac{q_0^{N_0}q_1^{N_1}q_2^{N_2}}{q^M(1-q)^{2n-M}}, \end{aligned}$$

où la plupart des termes peuvent être simplifiés. Cette probabilité est alors décrite par :

$$Q(H_{-1} \mid H_0) = \frac{\prod_s(m_s!)}{2^{N_1}(2n)!}\left[\frac{q_0^{N_0}q_1^{N_1}q_2^{N_2}}{q^M(1-q)^{2n-M}}\right]. \tag{5.5}$$

Notons que l'expression (5.5) généralise l'équation (5.4) que nous avions obtenu dans le cas d'un échantillon aléatoire simple. Dans ce cas particulier, $q_0 = (1-p)^2$, $q_1 = 2p(1-p)$, $q_2 = p^2$ et $q = p$. On retrouve alors

$$\begin{aligned} Q(H_{-1} \mid H_0) &= \frac{\prod_s (m_s!)}{2^{N_1}(2n)!}\left[\frac{(1-p)^{2N_0}(2p(1-p))^{N_1}p^{2N_2}}{p^M(1-p)^{2n-N}}\right] \\ &= \frac{\prod_s (m_s!)}{2^{N_1}(2n)!}\left[\frac{2^{N_1}(1-p)^{2N_0+N_1}(p)^{2N_2+N_1}}{p^M(1-p)^{2n-M}}\right] \\ &= \frac{\prod_s (m_s!)}{(2n)!}. \end{aligned}$$

5.1.5 Algorithme MapArg sur les génotypes

Nous venons de décrire comment il est possible d'intégrer la réalité diploïde à la méthode MapArg en ajoutant deux étapes au modèle d'échantillonnage pondéré. L'algorithme suivant résume les étapes de la nouvelle méthode.

Algorithme MapArg sur les génotypes

1. Choisir un ensemble fini de valeurs candidates r_T pour lesquelles la vraisemblance sera évaluée. Notons que la vraisemblance calculée n'est pas définie aux marqueurs de référence.
2. Estimer les distributions V_0 et V_1 par l'algorithme EM conditionnel aux génotypes décrit au chapitre 4.
3. Pour chaque intervalle m compris entre deux marqueurs de référence :
   (a) Pour chacun des K graphes à construire :
      i. Générer l'état H_{-1} en utilisant la distribution conditionnelle décrite à l'équation (4.14), page 67.
      ii. Construire la liste H_0 des séquences observées en insérant le marqueur de la mutation causale au centre de l'intervalle, de sorte qu'il se retrouve au rang m, tel qu'illustré à la figure 2.1 de la page 19. La position centrale r_{T_0} servira de valeur conductrice.

iii. Calculer la valeur de

$$\begin{aligned} h(H_{-2}, H_{-1}) &= \frac{Q(H_{-2} \mid H_{-1})}{P(H_{-1} \mid H_{-2})}, \\ h(H_{-1}, H_0) &= \frac{Q(H_{-1} \mid H_0)}{P(H_0 \mid H_{-1})} \end{aligned}$$

en utilisant les équations (5.1), (5.2), (5.3) et (5.5) développées à la section 5.1.2.

iv. Pour chaque étape τ telle que $0 \leq \tau \leq \tau^* - 1$:

A. Déterminer les états $H_{\tau+1}$ admissibles en considérant tous les événements possibles, tels que décrits à la section 2.2.1, page 21.

B. Pour chaque état, déterminer la probabilité associée en évaluant $b(H_\tau, H_{\tau+1})/f(H_\tau)$ (voir section 2.2.3, page 25).

C. Générer le nouvel état $H_{\tau+1}$ selon la distribution calculée à l'étape précédente.

D. Pour chaque valeur de r_T comprise dans l'intervalle, évaluer la fonction $h_{r_T r_{T_0}}(H_\tau, H_{\tau+1})$ décrite à l'équation (2.2), page 25.

v. Pour chaque valeur de r_T, calculer le produit $\prod_{\tau=-2}^{\tau^*-1} h_{r_T, r_{T_0}}(H_\tau, H_{\tau+1})$.

(b) Pour chaque position r_T de l'intervalle, calculer la moyenne sur tous les graphes :

$$\hat{Q}_{r_T}(H_{-2}) = (1/K) \sum_{k=1}^{K} \left[\prod_{\tau=-2}^{\tau^*-1} h_{r_T, r_{T_0}}(H_\tau^k, H_{\tau+1}^k) \right].$$

5.2 Vraisemblance sur les distributions V_0 et V_1

5.2.1 MapArg revisité

Nous venons de voir comment il est possible d'augmenter le modèle d'échantillonnage pondéré afin d'estimer la vraisemblance sur les génotypes et les phénotypes des individus. D'un point de vue théorique, cette façon de faire intègre la réalité diploïde dans son ensemble. Toutefois, celle-ci est largement dépendante du modèle de pénétrance,

qui n'est habituellement pas connu. Si l'évaluation de ce dernier n'est pas suffisamment précise, l'estimation obtenue risque d'être largement biaisée. Enfin, l'ajout de deux étapes lors de l'estimation rend les calculs plus laborieux. Une façon de contourner ces difficultés consiste à ramener le problème à une estimation sur des distributions haploïdes.

Rappelons que la méthode de cartographie que nous cherchons à généraliser nécessite un échantillon de séquences haploïdes pour lesquelles l'allèle au gène causal est supposé connu. Étant donné que les mutations étudiées sont rares, les échantillons utilisés comprennent habituellement une proportion fixée de séquences porteuses. Ainsi, on forme H_0 en choisissant aléatoirement n_1 séquences présentant la mutation causale, contre n_0 ne la présentant pas. Selon les notations que nous avons utilisé précédemment, ceci équivaut à échantillonner n_1 séquences de la distribution V_1, et n_0 de la distribution V_0. Ceci revient à dire que la version originale de MapArg suppose un échantillon provenant directement de V_0 et V_1.

5.2.2 Rééchantillonnage

La généralisation de MapArg à des individus diploïdes serait triviale si l'on pouvait échantillonner directement de V_0 et V_1. Malheureusement, ces distributions ne sont pas connues, et il n'y a aucun moyen d'en obtenir un échantillon de manière directe. La solution que nous proposons consiste donc à utiliser l'information sur les individus diploïdes afin d'estimer ces distributions, par l'algorithme EM conditionnel aux phénotypes développé au chapitre 4. Ainsi, l'unique différence par rapport à la version originale réside dans le fait que l'échantillon H_0 est obtenu en échantillonnant de $\hat{V}_0$ et $\hat{V}_1$. Ainsi, l'optimisation de $Q(H_0 \mid r_T)$ visera maintenant à maximiser $Q(\hat{V}_0, \hat{V}_1 \mid r_T)$. Rappelons que les distributions $\hat{V}_0$ et $\hat{V}_1$ obtenues par l'algorithme EM sont telles que $Q(G, \Phi \mid V_0, V_1)$ est optimisée. De ce fait, la méthode proposée équivaut approximativement à maximiser $Q(G, \Phi \mid r_T)$.

La méthode que nous proposons ici a l'avantage d'être simple et beaucoup plus robuste à une mauvaise estimation des paramètres f_0, f_1 et f_2 que la vraisemblance sur les données observées. En effet, l'estimation de V_0 et V_1 est beaucoup moins sensible au modèle de pénétrance que le calcul de la vraisemblance. Évidemment, nous sommes conscients que l'application de cette méthode nécessite des compromis d'un point de vue idéologique, puisque la vraisemblance n'est pas calculée sur un échantillon véritablement tiré d'une population. Toutefois, celle-ci pourrait être apparentée à des techniques de rééchantillonnage.

5.3 Vraisemblance composite revisitée

5.3.1 Vraisemblance composite sur les génotypes

Nous avons vu au second chapitre comment les vraisemblances composites pouvaient aider à réduire le temps de calcul. Appliquons maintenant ce même raisonnement pour des individus diploïdes. Découpons l'information d'origine en J fenêtres de marqueurs, formant ainsi un ensemble d'événements $\{A_j, j = 1, .., J\}$, où A_j représente les marqueurs et intervalles de la fenêtre j. L'application directe de l'équation (2.3), page 27, permet d'écrire la vraisemblance composite :

$$\begin{aligned} CL(r_T) &= \prod_j Q(G, \Phi, A_j \mid r_T) \\ &= \prod_j Q(H_{-2}, A_j \mid r_T). \end{aligned}$$

Reprenons les découpages décrits au second chapitre. Considérons dans un premier temps des fenêtres juxtaposées (figure 2.3, page 28). Supposons que la mutation causale, située dans l'intervalle m, est incluse dans la fenêtre j_m. Remarquons que, pour tout $j \neq j_m$,

$$Q(H_{-2}, A_j \mid r_T) = Q(H^*_{-2}, A_j),$$

où $H^*_{-2} = G$ représente l'information sur les génotypes aux marqueurs de référence uniquement. En effet, les fenêtres ne comprenant pas l'intervalle m ne contiennent aucune

information sur la mutation causale et, par le fait même, sur le phénotype. On peut alors décrire, pour chaque intervalle m, la vraisemblance composite définie sur celui-ci par :

$$CL_m(r_T) = \frac{Q(H_{-2}, A_{j_m} \mid r_T)}{Q(H^*_{-2}, A_{j_m})} \prod_j Q(H^*_{-2}, A_j).$$

Puisque le facteur $\prod_j Q(H^*_{-2}, A_j)$ de l'expression précédente ne dépend que du découpage en fenêtres, toute l'information sur la position r_T est contenue dans le quotient. Définissons alors, pour chaque fenêtre j et chaque intervalle m, la fonction

$$L_{m,j}(r_T \mid H^*_{-2}, A_{j_m}) = \begin{cases} \frac{Q(H_{-2}, A_{j_m} \mid r_T)}{Q(H^*_{-2}, A_{j_m})}, & \text{si } r_T \text{ est inclus dans l'intervalle } m, \text{ fenêtre } j \\ 1, & \text{sinon.} \end{cases}$$

Comme précédemment, on retrouve la vraisemblance composite conditionnelle sur l'ensemble de la séquence en effectuant le produit des vraisemblances conditionnelles sur chaque intervalle :

$$CCL(r_T) = \prod_{m=2}^{L-1} L_{m,j_m}(r_T \mid H^*_{-2}, A_{j_m}).$$

Généralisons ce que nous venons de décrire. Considérons maintenant des fenêtres superposées de d marqueurs, de sorte qu'il y ait un pas de un marqueur entre chaque fenêtre (figure 2.4, page 30). Rappelons que l'intervalle m est inclus dans toutes les fenêtres g telles que $\underline{j}(m) \leq g \leq \overline{g}(m)$ où

$$\begin{aligned} \underline{j}(m) &= max(1, m+1-d), \\ \overline{j}(m) &= min(m-1, L-d). \end{aligned}$$

Cet intervalle est donc inclus dans $\overline{j}(m) - \underline{j}(m) + 1$ fenêtres. En effectuant la moyenne géométrique, on trouve la vraisemblance composite conditionnelle

$$CCL_d(r_T) = \prod_{m=2}^{L-1} \left(\prod_{j=\underline{j}(m)}^{\overline{j}(m)} L_{m,j}(r_T \mid H^*_{-2}, A_j) \right)^{\omega_m},$$

où

$$\omega_m = \frac{1}{\overline{j}(m) - \underline{j}(m) + 1}.$$

Cette vraisemblance composite conditionnelle est une approximation de la vraisemblance sur les données complètes, à un facteur $\prod_j Q(H^*_{-2}, A_j)$ près dépendant du découpage choisi. Notons que le calcul de $Q(H^*_{-2})$ se fait de manière similaire à celui de $Q(H_{-2})$. Il suffit de ne pas tenir compte du phénotype et de la mutation causale dans les calculs de probabilités.

5.3.2 Vraisemblance composite sur H_0

Nous avons vu précédemment comment, en principe, il est possible d'obtenir un échantillon H_0 à partir des distributions $\hat{V}_0$ et $\hat{V}_1$ et de l'utiliser directement dans MapArg. En pratique, le nombre de marqueurs est limité par l'algorithme EM utilisé. De nouveau, la vraisemblance composite permet l'application à plus grande échelle.

Reprenons le découpage en J fenêtres superposées de marqueurs (figure 2.4, page 30). Considérons maintenant un découpage de H_0 en autant de sous-échantillons comprenant chacun m_1 et m_0 séquences respectivement porteuses et non porteuses. Afin de ne pas biaiser l'estimation de la vraisemblance, l'attribution des séquences doit s'effectuer au hasard et les valeurs de m_0 et m_1 doivent être constantes. Associons à la fenêtre j le sous-échantillon j, pour $j = 1, ..., J$. Ainsi, on obtient un nouvel ensemble d'événements $\{A_j, 1 \leq j \leq J\}$, où A_j représente l'information sur les marqueurs et intervalles de la fenêtre et du sous-échantillon j.

La vraisemblance composite se calcule alors exactement de la même façon que lorsque les événements ne comportaient que des fenêtres de marqueurs. L'équation (2.5), à la page 30, peut alors être appliquée directement.

Cette façon de construire la vraisemblance composite équivaut en pratique à générer un sous-échantillon de séquences pour chaque fenêtre j. Pour ce faire, on obtient les distributions $\hat{V}_0^j$ et $\hat{V}_1^j$ en appliquant l'algorithme EM généralisé sur la fenêtre de marqueurs correspondante. Il suffit alors d'échantillonner le nombre déterminé d'éléments selon ces distributions.

Nous avons développé dans ce chapitre deux méthodes permettant d'intégrer la réalité diploïde et les modèles de pénétrance à la méthode Maparg. Celles-ci font appel à l'algorithme EM conditionnel aux phénotypes décrit au chapitre 4. Au prochain chapitre, nous appliquerons ces algorithmes à des données simulées.

CHAPITRE VI

RÉSULTATS

Dans les chapitres précédents, nous avons discuté de la problématique de l'estimation des haplotypes et du statut au gène causal, dans le contexte de la cartographie génétique. Dans l'optique d'y apporter une solution, nous avons décrit au chapitre 4 un algorithme EM conditionnel au phénotype pouvant être appliqué à des échantillons à proportion fixée de cas. Nous avons également expliqué au chapitre 5 comment cet algorithme pouvait être utilisé afin d'intégrer la réalité diploïde à la méthode de cartographie génétique MapArg. L'objectif de ce chapitre est maintenant d'évaluer la performance de l'algorithme EM et l'impact de son utilisation dans le contexte de la méthode MapArg. Nous tenterons de déterminer les limites d'application de l'algorithme, ainsi que les conditions optimales. Pour ce faire, nous considérerons dans un premier temps des modèles de pénétrance connus. Nous traiterons par la suite le cas de modèles inconnus.

6.1 Distributions estimées

6.1.1 Présentation et méthodologie

Nous avons développé au chapitre 4 un algorithme EM conditionnel aux phénotypes nous permettant d'estimer les distributions V_0 et V_1, respectivement associées aux séquences non porteuses et porteuses de la population. L'objectif de cette section est d'en évaluer la performance. Pour ce faire, nous en avons programmé une version en langage C++ que nous avons utilisée pour effectuer de nombreuses simulations.

Les échantillons utilisés pour nos tests ont été obtenus de populations haploïdes simulées selon le principe de coalescence. Ces populations de 10 000 haplotypes chacune ont été simulées à l'aide du programme *ms* de Hudson (2002). Le taux global de recombinaison a été fixé à $\rho = 100$. Plus de détails sur la constructions des populations haploïdes sont donnés dans l'article de Larribe et Lessard (2008). Nous avons choisi six de ces populations, identifiées par les lettres A, B, C, D, E et F. Pour chacune d'elles, un des marqueurs a été choisi comme mutation causale. Celui-ci a été sélectionné de manière à atteindre une fréquence approximative de 10%. Nous avons ensuite généré une population diploïde en jumelant les haplotypes de manière aléatoire. Pour chaque individu, le phénotype a été généré selon le modèle de pénétrance choisi. Un échantillon à proportion fixée de 50 cas et 50 témoins a finalement été tiré de la population diploïde. Nous avons obtenu deux séries de données pour chaque échantillon. Dans la première série, le marqueur de la mutation a été retiré, et les haplotypes mêlés pour former des génotypes. Construit de cette façon, cet échantillon correspond à ce qui est généralement obtenu en laboratoire. La seconde série contient quant à elle les séquences haploïdes associées aux individus, incluant le gène causal. Évidemment, ce type d'échantillon n'est pas observable en pratique et correspond en quelque sorte à la solution que l'on souhaite atteindre par l'algorithme EM.

Afin d'avoir une vision générale de la performance de l'algorithme, quelques modèles différents ont été testés. Nous avons choisi un modèle récessif parfait, un modèle récessif avec 1% de phénocopies, un modèle dominant avec 1% de phénocopies, et un modèle mixte avec 5% de phénocopies. Rappelons que le taux de phénocopies est la proportion d'individus non porteurs qui présentent le caractère d'intérêt. Dans le contexte d'une maladie rare, la proportion de faux positifs peut devenir très élevée, malgré un faible taux de phénocopies. Par exemple, pour le modèle mixte décrit ici, les phénocopies représentent plus de 60% des individus cas de la population, contre seulement 12% pour les doublement porteurs. Pour le modèle mixte, nous avons donc généré un second échantillon de plus grande taille (environ 100 cas et 100 témoins). Le tableau 6.1 décrit les divers échantillons utilisés.

Tableau 6.1 Description des échantillons*

Modèle / Population	A	B	C	D	E	F
$f_0 = 0,00;\ f_1 = 0,00;\ f_2 = 1,00$	Ap	Bp	Cp	Dp	Ep	Fp
$f_0 = 0,01;\ f_1 = 0,01;\ f_2 = 0,95$	Ar	Br	Cr	Dr	Er	Fr
$f_0 = 0,01;\ f_1 = 0,90;\ f_2 = 0,95$	Ad	Bd	Cd	Dd	Ed	Fd
$f_0 = 0,05;\ f_1 = 0,10;\ f_2 = 0,80$	Am	Bm	Cm	Dm	Em	Fm
	Am2	Bm2	Cm2	Dm2	Em2	Fm2

*. Les lettres majuscules correspondent aux populations. Chaque modèle est identifié par une lettre minuscule : «p» pour récessif parfait, «r» pour récessif avec phénocopies, «d» pour dominant et «m» pour mixte. Le «2» identifie l'échantillon mixte de plus grande taille.

La performance de l'algorithme EM conditionnel aux phénotypes a été évaluée en comparant les distributions V_0 et V_1 estimées par ce dernier sur les échantillons diploïdes avec celles obtenues directement des séquences haploïdes correspondantes. Cette comparaison a été faite à l'aide de graphiques en barres associés aux distributions. Chaque haplotype compatible avec les génotypes observés y est représenté par une barre dont la longueur est proportionnelle à la fréquence de celui-ci. Notons que, les séquences non compatibles ayant été retirées, les paramètres estimés nuls sont informatifs sur la performance de l'algorithme. Le graphique obtenu dresse un profil de la distribution. Afin de mieux comparer ces profils, les distributions estimées et celles observées ont été superposées sur un même graphique.

6.1.2 Application du modèle exact

L'algorithme EM que nous avons présenté nécessite l'estimation d'un grand nombre de paramètres. Rappelons que la dimension des distributions à estimer augmente de façon exponentielle avec la taille de la fenêtre. Chaque marqueur présentant deux allèles, il existe théoriquement 2^d haplotypes possibles pour une séquence de taille d. Puisque les paramètres doivent sommer à un, il en reste $2^d - 1$ à estimer pour chacune des deux

distributions. Ainsi, une fenêtre de 4 marqueurs nécessitera l'estimation de $2^4 - 1 = 15$ paramètres par distribution, pour un total de 30. Pour des fenêtres de taille 6, 8 ou 10, il faudra en estimer respectivement 70, 126 ou 198. Notons toutefois que tous les haplotypes théoriques ne sont pas nécessairement compatibles avec les génotypes observés, ce qui réduit la taille effective des distributions.

Pour un échantillon d'une centaine d'individus, il est raisonnable de penser que des fenêtres de 4 à 8 marqueurs soient appropriées. Les figures 6.1, 6.2 et 6.3 illustrent des résultats pour des fenêtres de 4, 6 et 8 marqueurs respectivement. Pour ces trois comparaisons, la fenêtre a été choisie de manière à être centrée sur la vraie position de la mutation. Plus le nombre de marqueurs augmente, plus il y a d'erreurs dans l'estimation. Nous avons remarqué que les erreurs se font habituellement d'une distribution vers une autre. Ainsi, plusieurs erreurs n'impliquent pas de mauvais haplotypes, mais plutôt une attribution de ceux-ci à la mauvaise population haploïde. Par exemple, à la figure 6.2, la fréquence du troisième haplotype en partant de la droite est surestimée parmi les séquences porteuses et sous-estimée parmi les non porteuses. Nos tests ne nous ont pas permis de constater une amélioration de l'estimation lorsque la taille de l'échantillon est doublée. Nous pensons qu'il faudrait une taille d'échantillon beaucoup plus grande pour constater une amélioration. Par exemple, la figure 6.4 a été obtenue sur une fenêtre de 6 marqueurs à partir d'un échantillon de 200 individus, soit le double de celui qui a permis de construire la figure 6.2.

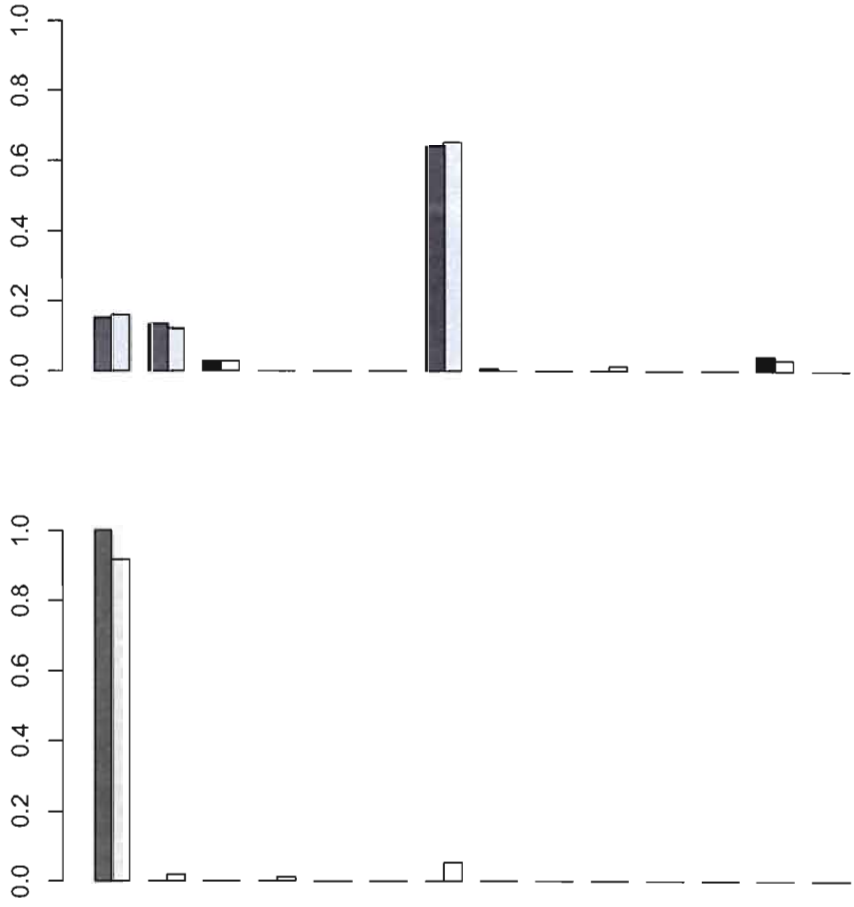

Figure 6.1 Distributions V_0 et V_1 sur les données Bm (fenêtre de quatre marqueurs, centrée). En haut, la distribution V_0. En bas, la distribution V_1. La distribution estimée par algorithme EM est illustrée en gris pâle, tandis que la distribution obtenue sur les données haploïdes est en gris foncé. Chaque barre représente une séquence compatible avec les génotypes observés.

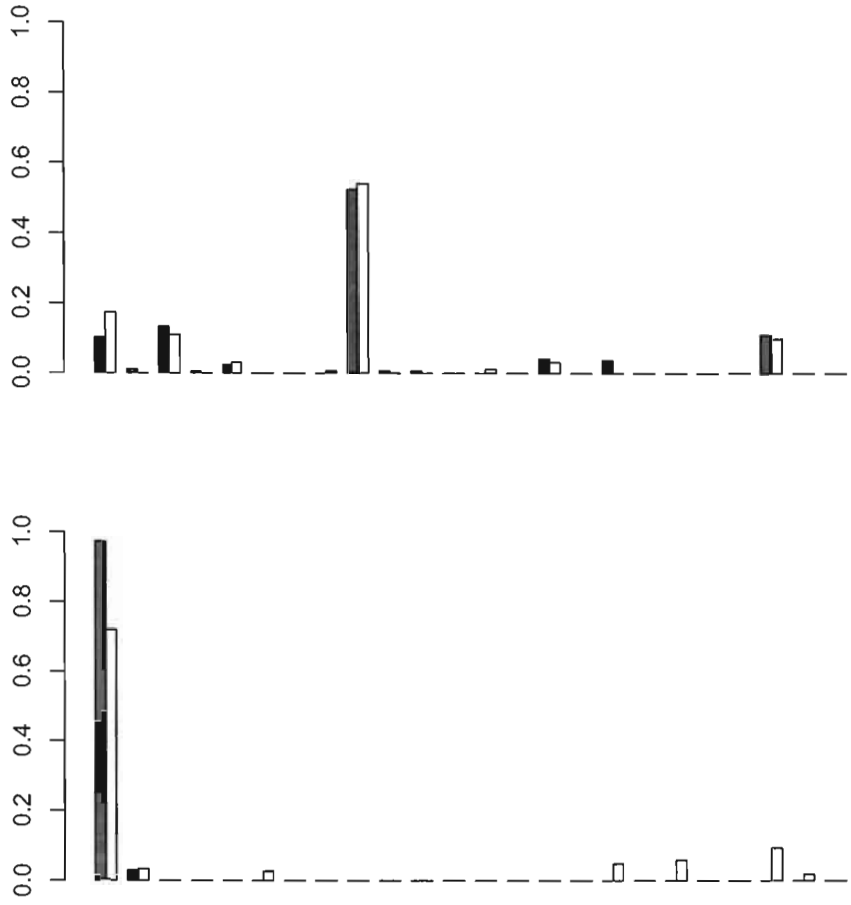

Figure 6.2 Distributions V_0 et V_1 sur les données Bm (fenêtre de six marqueurs, centrée). En haut, la distribution V_0. En bas, la distribution V_1. La distribution estimée par algorithme EM est illustrée en gris pâle, tandis que la distribution obtenue sur les données haploïdes est en gris foncé. Chaque barre représente une séquence compatible avec les génotypes observés.

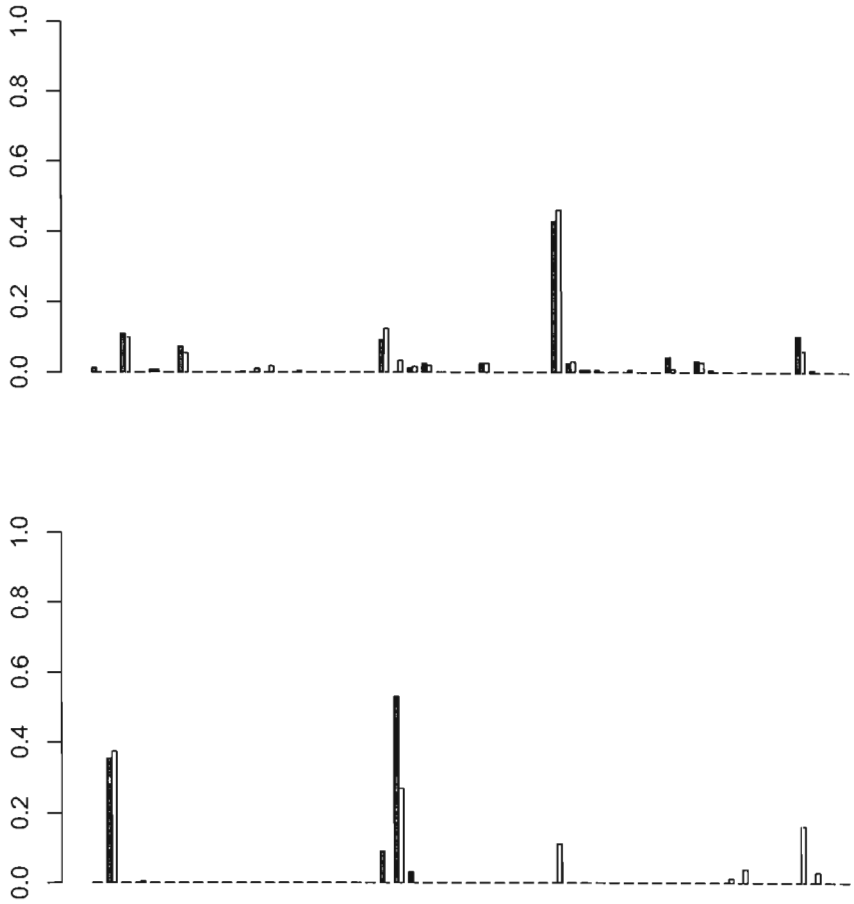

Figure 6.3 Distributions V_0 et V_1 sur les données Bm (fenêtre de huit marqueurs, centrée). En haut, la distribution V_0. En bas, la distribution V_1. La distribution estimée par algorithme EM est illustrée en gris pâle, tandis que la distribution obtenue sur les données haploïdes est en gris foncé. Chaque barre représente une séquence compatible avec les génotypes observés.

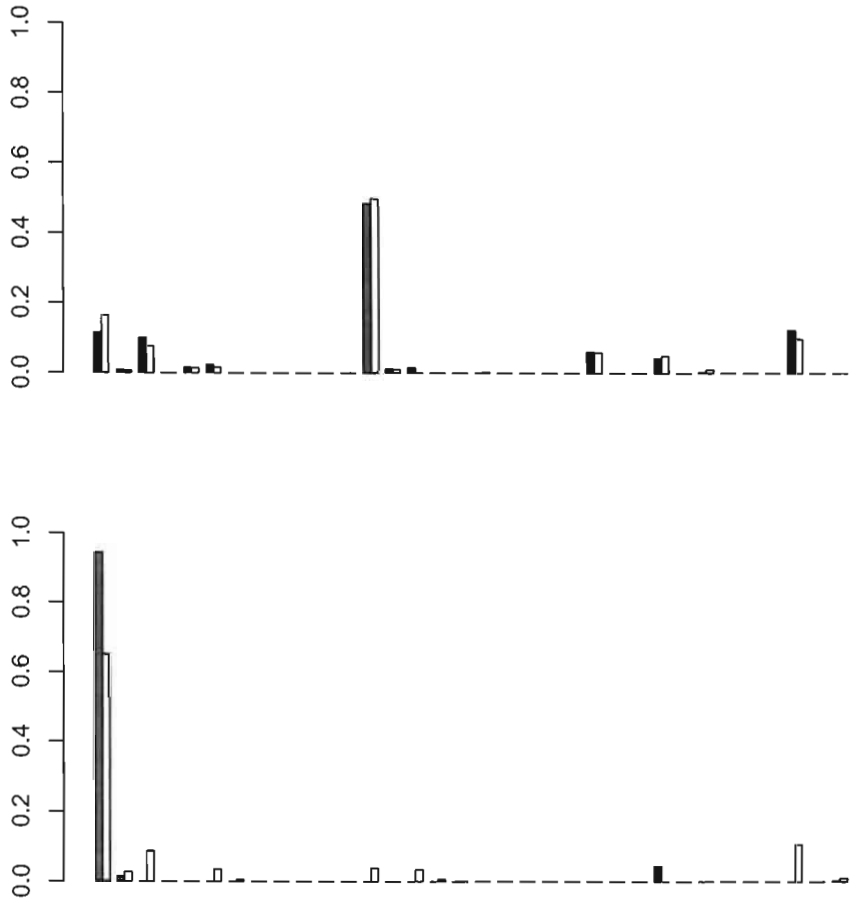

Figure 6.4 Distributions V_0 et V_1 sur les données Bm2 (fenêtre de six marqueurs, centrée). En haut, la distribution V_0. En bas, la distribution V_1. La distribution estimée par algorithme EM est illustrée en gris pâle, tandis que la distribution obtenue sur les données haploïdes est en gris foncé. Chaque barre représente une séquence compatible avec les génotypes observés.

La méthode MapArg a été conçue dans l'objectif d'estimer la position de mutations causales relativement rares. Pour cette raison, un échantillon aléatoire simple présentera très peu de séquences porteuses, donc peu d'information sur la distribution V_1. Nous avons vu au chapitre 4 que l'algorithme EM conditionnel aux phénotypes pouvait être appliqué à un échantillon à proportion fixée de cas. Lorsqu'il y a peu de phénocopies, c'est-à-dire peu d'individus cas non porteurs, ce type d'échantillon contient une proportion importante de séquences porteuses. D'un autre côté, si l'échantillon contient des phénocopies, il peut y avoir une très grande proportion d'individus non porteurs présentant le caractère d'intérêt. Pour cette raison, l'échantillon recèle parfois très peu d'information sur V_1, ce qui peut affecter l'estimation de celle-ci par l'algorithme EM.

Plus le modèle présente des phénocopies, plus il est difficile d'estimer la distribution V_1. En observant les figures 6.1, 6.2 et 6.3, on remarque que l'estimation de V_0 est généralement meilleure que celle de V_1. L'échantillon utilisé dans ces exemples comporte un nombre non négligeable de phénocopies. Ainsi, sur 50 individus cas, 30 sont des phénocopies, 15 sont simplement porteurs et 5 possèdent deux copies du gène. On note que l'estimation de V_1 est nettement améliorée lorsque le nombre de phénocopies est diminué dans l'échantillon. Par exemple, l'estimation de V_1 est presque exacte dans la figure 6.5. Ce résultat a été obtenu sur un échantillon auquel on a appliqué un modèle récessif ne présentant aucune phénocopie. Dans ce cas, les 50 individus cas sont doublement porteurs. La figure 6.6 donne quant à elle un exemple d'estimation obtenue avec un modèle dominant présentant 1% de phénocopies. Parmi les 50 individus cas de cet échantillon, seulement 2 sont des phénocopies, 4 sont doublement porteurs et 44 possèdent une unique copie de la mutation.

Plus on s'éloigne de la mutation, plus on s'attend à ce que les distributions V_0 et V_1 se ressemblent. En effet, les marqueurs situés à proximité de la mutation causale sont plus souvent transmis avec elle, ce qui est à l'origine du déséquilibre de liaison. De ce fait, on s'attend aussi à une perte graduelle de l'information sur V_1 au fur et à mesure que l'on s'éloigne du gène causal. Les tests que nous avons effectués confirment généralement ces hypothèses. Pour plusieurs échantillons, plus on s'éloigne de la mutation causale, plus les

profils des distributions V_0 et V_1 se ressemblent. De même, l'estimation de V_1 présente un peu plus d'erreurs. La figure 6.7 compare les distributions obtenues sur une fenêtre de 4 marqueurs très éloignée de la mutation causale. On remarque que la distribution V_0 est très bien estimée, tandis qu'il y a un légèrement plus d'erreur sur la distribution V_1.

Afin de comparer les vecteurs V_0 et V_1, nous avons calculé une statistique de distance géométrique entre ces deux vecteurs. Cette distance est simplement obtenue en sommant les carrés des différences aux coordonnées. Les figures 6.8 et 6.9 donnent ces distances en fonction de la position du centre de la fenêtre, pour 6 échantillons récessifs avec 1% de phénocopies. On remarque que cette seule statistique est parfois suffisante pour identifier la position du gène causal. D'ailleurs, le calcul de distance rappelle certaines méthodes actuellement utilisées en cartographie génétique. Nous avons inséré dans le même graphique les distances estimées par l'algorithme EM et celles obtenues directement des séquences haploïdes. Plus les courbes sont rapprochées, plus on peut supposer que l'algorithme EM a bien départagé les haplotypes porteurs et non porteurs.

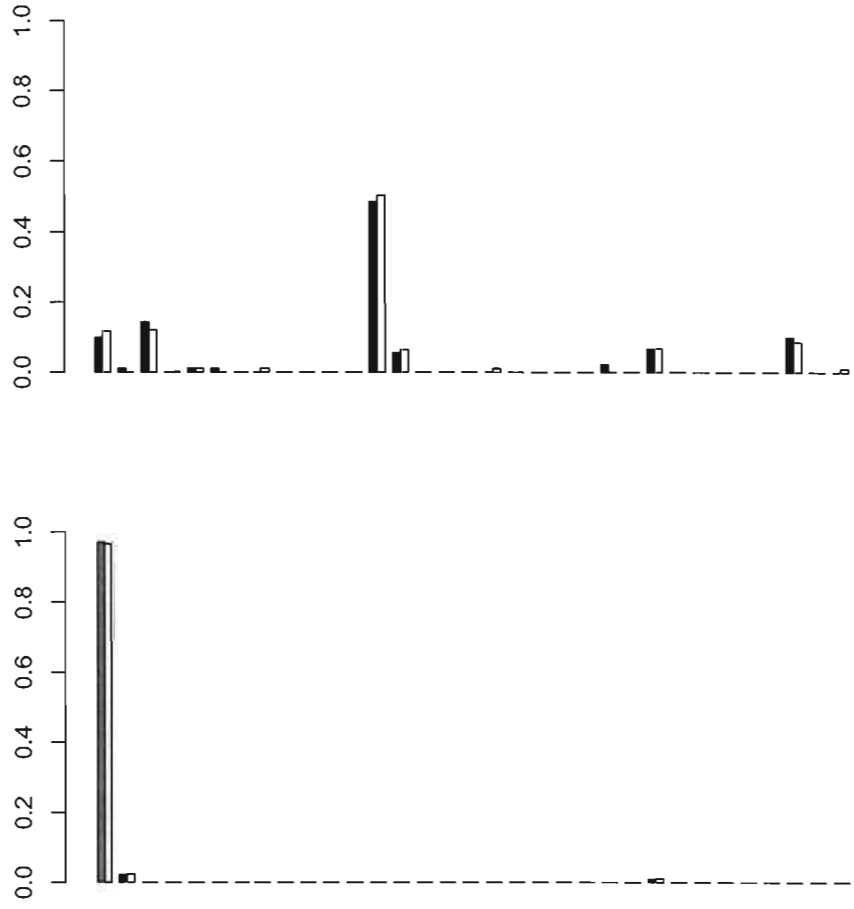

Figure 6.5 Distributions V_0 et V_1 sur les données Bp (fenêtre de six marqueurs, centrée). En haut, la distribution V_0. En bas, la distribution V_1. La distribution estimée par algorithme EM est illustrée en gris pâle, tandis que la distribution obtenue sur les données haploïdes est en gris foncé. Chaque barre représente une séquence compatible avec les génotypes observés.

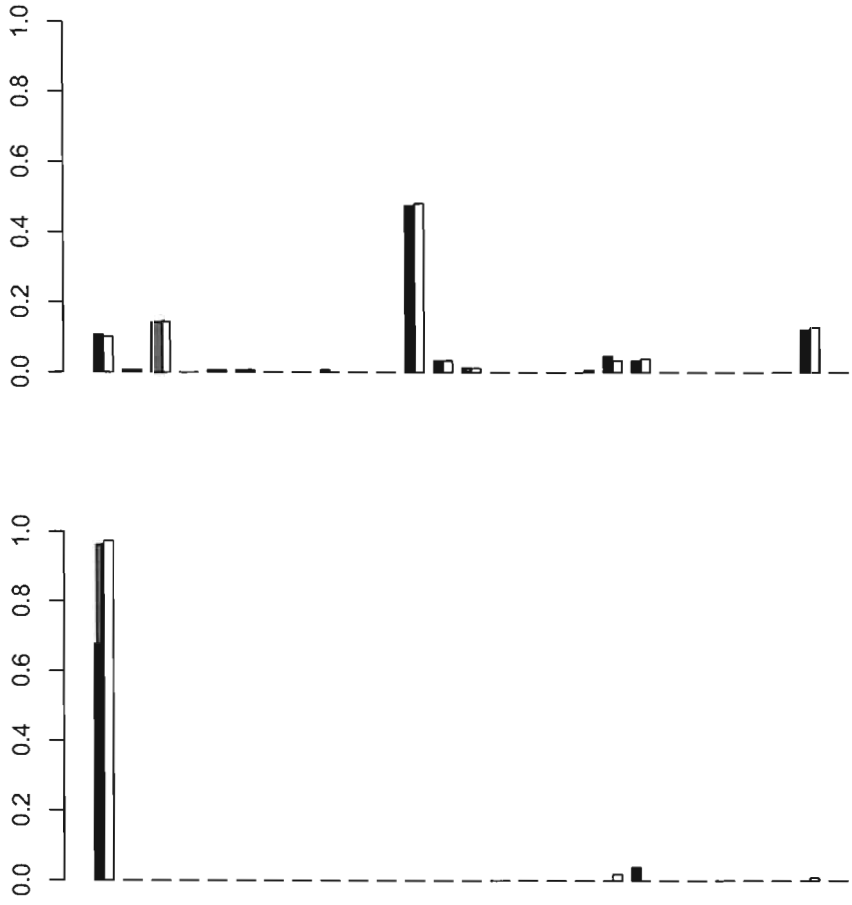

Figure 6.6 Distributions V_0 et V_1 sur les données Bd (fenêtre de six marqueurs, centrée). En haut, la distribution V_0. En bas, la distribution V_1. La distribution estimée par algorithme EM est illustrée en gris pâle, tandis que la distribution obtenue sur les données haploïdes est en gris foncé. Chaque barre représente une séquence compatible avec les génotypes observés.

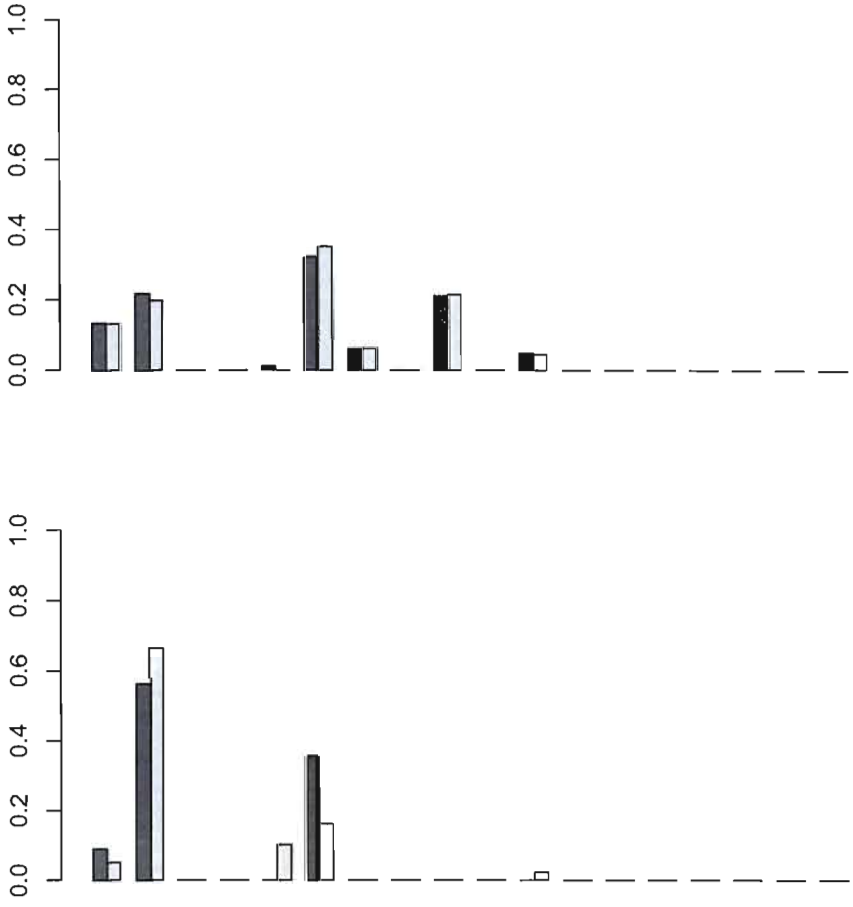

Figure 6.7 Distributions V_0 et V_1 sur les données Bm (fenêtre de quatre marqueurs, décentrée). En haut, la distribution V_0. En bas, la distribution V_1. La distribution estimée par algorithme EM est illustrée en gris pâle, tandis que la distribution obtenue sur les données haploïdes est en gris foncé. Chaque barre représente une séquence compatible avec les génotypes observés.

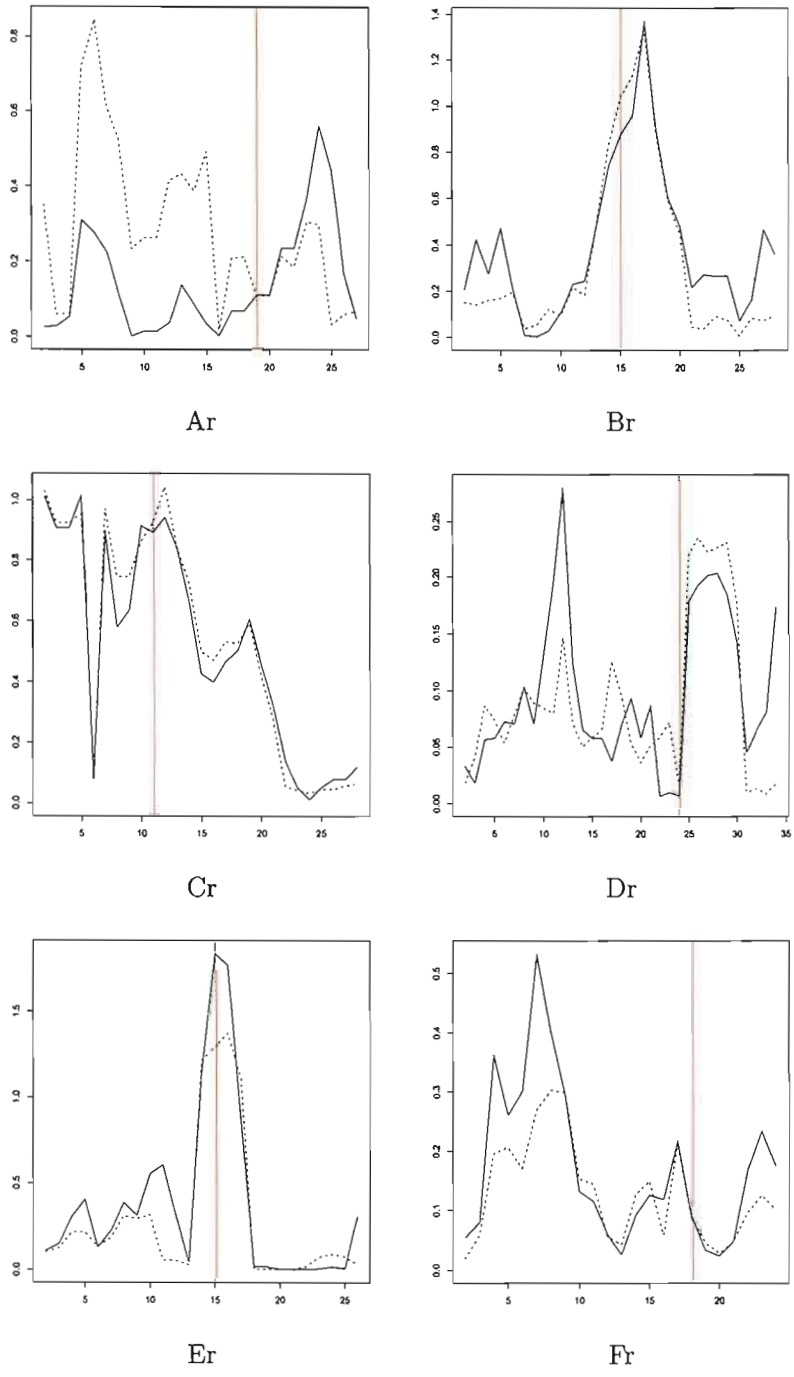

Figure 6.8 Graphiques de distances entre V_0 et V_1. La courbe pleine représente la distance estimée par l'algorithme EM. La courbe en pointillés correspond à la distance obtenue directement des séquences. La ligne verticale est la position réelle de la mutation. Pour ces exemples, des fenêtres de 4 marqueurs ont été utilisées.

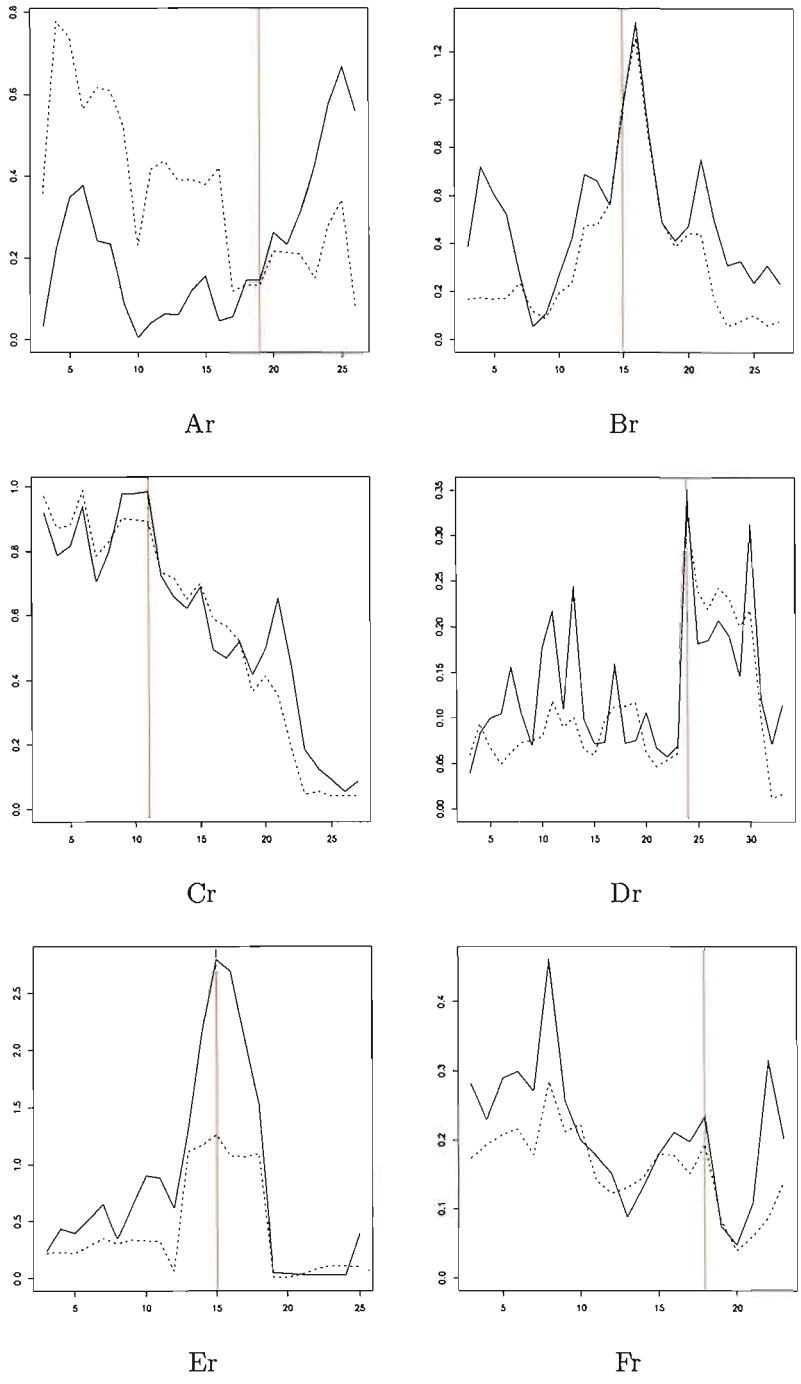

Figure 6.9 Graphiques de distances entre V_0 et V_1. La courbe pleine représente la distance estimée par l'algorithme EM. La courbe en pointillés correspond à la distance obtenue directement des séquences. La ligne verticale est la position réelle de la mutation. Pour ces exemples, des fenêtres de 6 marqueurs ont été utilisées.

6.2 MapArg et réalité diploïde

6.2.1 Intégration arbitraire

Nous avons décrit au chapitre 2 une méthode de cartographie génétique basée sur le graphe de recombinaison ancestral. Pour appliquer celle-ci, un échantillon de séquences haploïdes doit être disponible. De plus, le statut au gène causal doit être connu, c'est-à-dire que l'on doit avoir préalablement identifié les séquences porteuses de la mutation causale. Nous avons mentionné que cette hypothèse est peu réaliste en pratique. En effet, les données recueillies prennent habituellement la forme de génotypes et de phénotypes. Une solution simple et arbitraire peut être utilisée pour pallier à ce problème. Il est possible d'estimer les haplotypes de manière tout à fait aléatoire parmi les combinaisons possibles, puis de définir le statut au gène causal directement à partir du phénotype. Ainsi, on supposera porteuse de la mutation toute séquence appartenant à un individu cas. Les haplotypes seront quant à eux générés aléatoirement en attribuant une même probabilité à tous les diplotypes compatibles avec le génotype.

La solution que nous venons de décrire consiste en une intégration plutôt arbitraire de la réalité diploïde, puisqu'on ne tient alors aucunement compte de la complexité des traits observés. En effet, si la population présente des phénocopies, il est fort probable qu'un grand nombre d'individus cas ne portent aucune copie du gène causal. De plus, si le modèle est récessif, un grand nombre d'individus témoins porteront une copie de la mutation. La figure 6.10 illustre l'effet de cette méthode arbitraire sur l'estimation de la position de la mutation causale. On peut comparer ces résultats à ceux obtenus à partir des véritables séquences haploïdes tels qu'illustrés à la figure 6.11. Dans les deux cas, les fenêtres utilisées pour la vraisemblance composite comportent 6 marqueurs. On remarque pour la plupart des échantillons une perte d'information lorsque l'intégration de la réalité diploïde se fait de manière arbitraire.

Rappelons que les vraisemblances estimées sont aléatoires, puisque la méthode de cartographie est elle-même basée sur un modèle d'échantillonnage pondéré. Pour tous nos graphiques, nous avons fait 10 essais indépendants. Pour chaque intervalle de chaque fenêtre, nous avons généré 500 graphes, ce qui représente de 50 000 à 100 000 graphes par essai, selon le nombre et la taille des fenêtres. La courbe bleue est une combinaison de ces 10 essais et est donc obtenue à partir de plus de 500 000 graphes générés de manière indépendante.

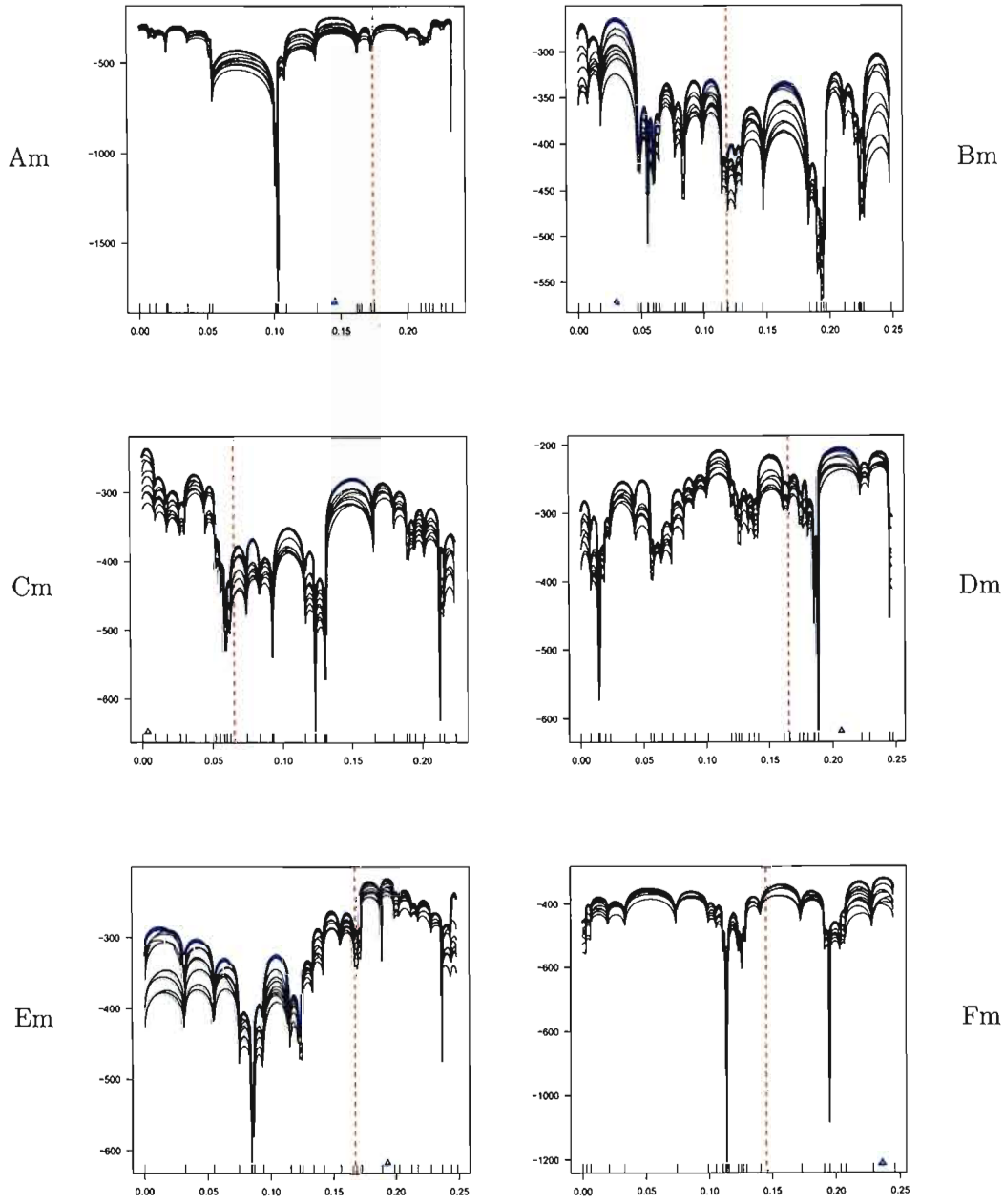

Figure 6.10 MapArg avec intégration arbitraire de la réalité diploïde. Chaque courbe représente un essai indépendent. La courbe en bleu est obtenue en combinant les essais. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

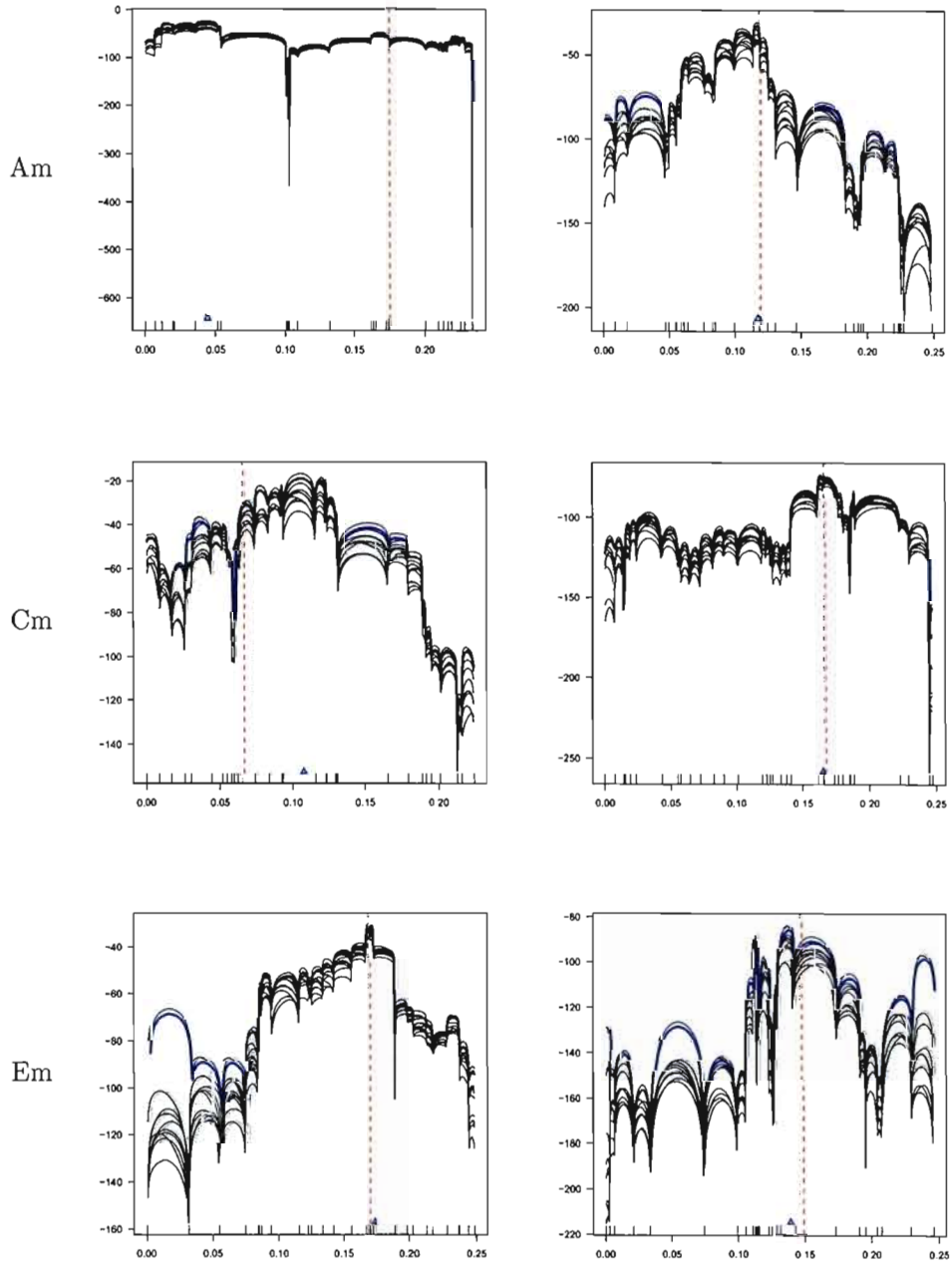

Figure 6.11 MapArg sur les données solution. Chaque courbe représente un essai indépendent. La courbe en bleu est obtenue en combinant les essais. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

6.2.2 Intégration de la réalité diploïde par algorithme EM

Nous avons décrit au chapitre 5 deux manières d'intégrer la réalité diploïde à la méthode MapArg. Toutes deux sont basées sur l'algorithme EM conditionnel aux phénotypes développé au chapitre 4. L'objectif de cette section est d'en évaluer les performances. Pour ce faire, nous avons effectué de nombreux tests de simulation avec différents échantillons et longueurs de fenêtres. Comme nous l'avons vu dans une section précédente, l'algorithme EM performe généralement bien pour des fenêtres de 4 à 6 marqueurs. Pour cette raison, nous avons choisi de limiter nos tests à ces dimensions. Aussi, les essais effectués sur les véritables séquences haploïdes nous ont permis de constater que, pour les populations A et C, ces tailles de fenêtres ne permettaient pas d'obtenir des estimations précises de la position de la mutation par la méthode MapArg, même avec une pénétrance complète et les haplotypes connus. Ainsi, nous ne reporterons dans cette section que les tests effectués sur les échantillons obtenus des populations B, D, E et F.

Afin de mieux évaluer les deux méthodes, nous avons choisi de regrouper les profils de vraisemblance obtenus à partir d'une même population. Ainsi, trois échantillons ont été testés pour chaque population, chacun correspondant à un modèle de pénétrance avec phénocopies : un dominant, un récessif, et un mixte. Les échantillons utilisés sont ceux décrits au tableau 6.1 de la page 103. Dans tous les cas, 10 essais indépendants ont été effectués, en générant 500 graphes pour chacun. Pour la méthode de rééchantillonnage, nous avons choisi de générer 500 séquences, dont 50 mutantes. Ce choix se justifie par la volonté d'utiliser le plus possible l'information sur les distributions estimées, tout en conservant un nombre raisonnable de séquences. Du même coup, on corrige un certain biais de sélection pouvant influencer la construction des graphes. Les figures 6.12 à 6.19 présentent les courbes de vraisemblance estimées.

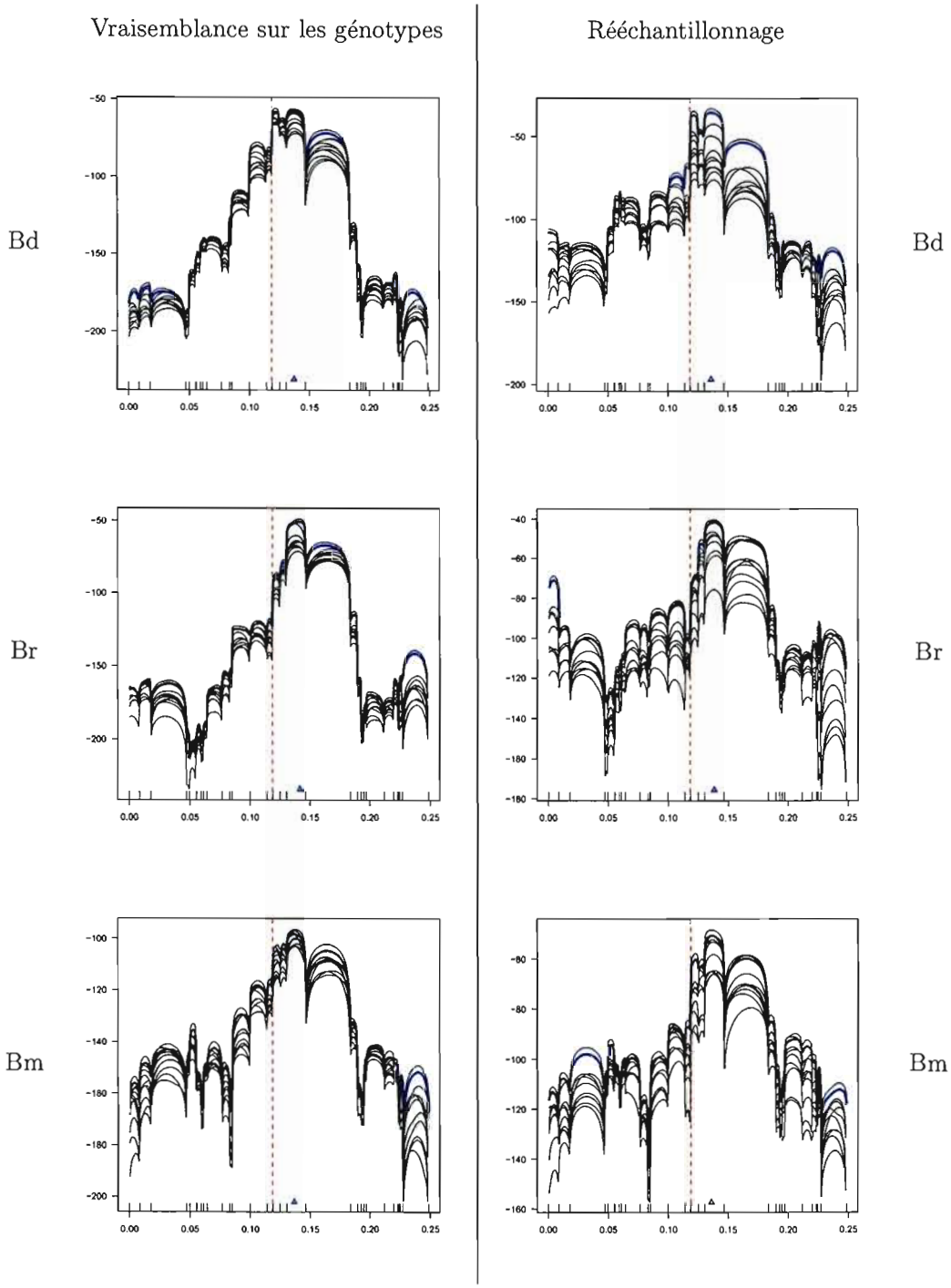

Figure 6.12 MapArg sur les données B avec des fenêtres de 4 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est supposé connu. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

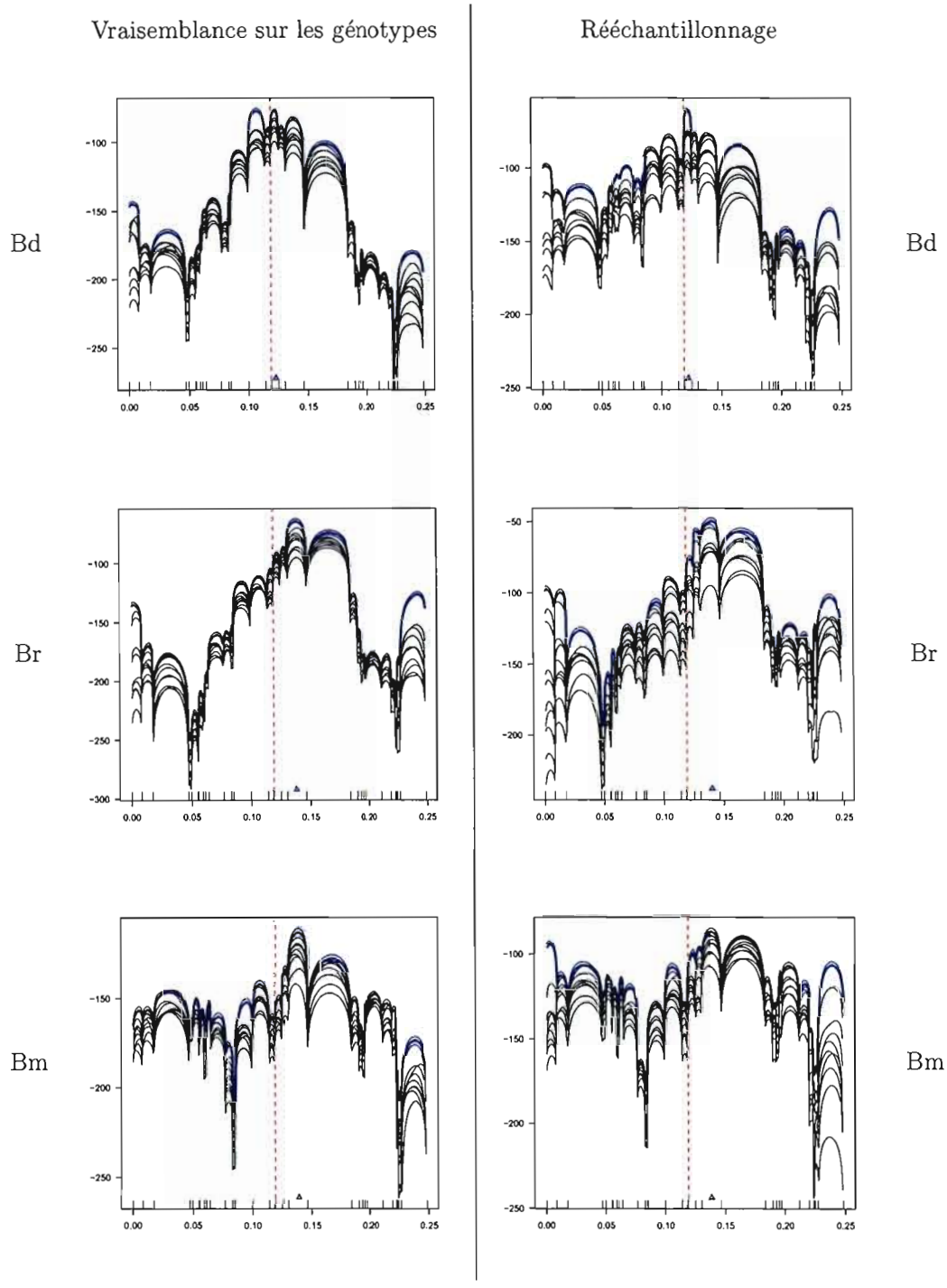

Figure 6.13 MapArg sur les données B avec des fenêtres de 6 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est supposé connu. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

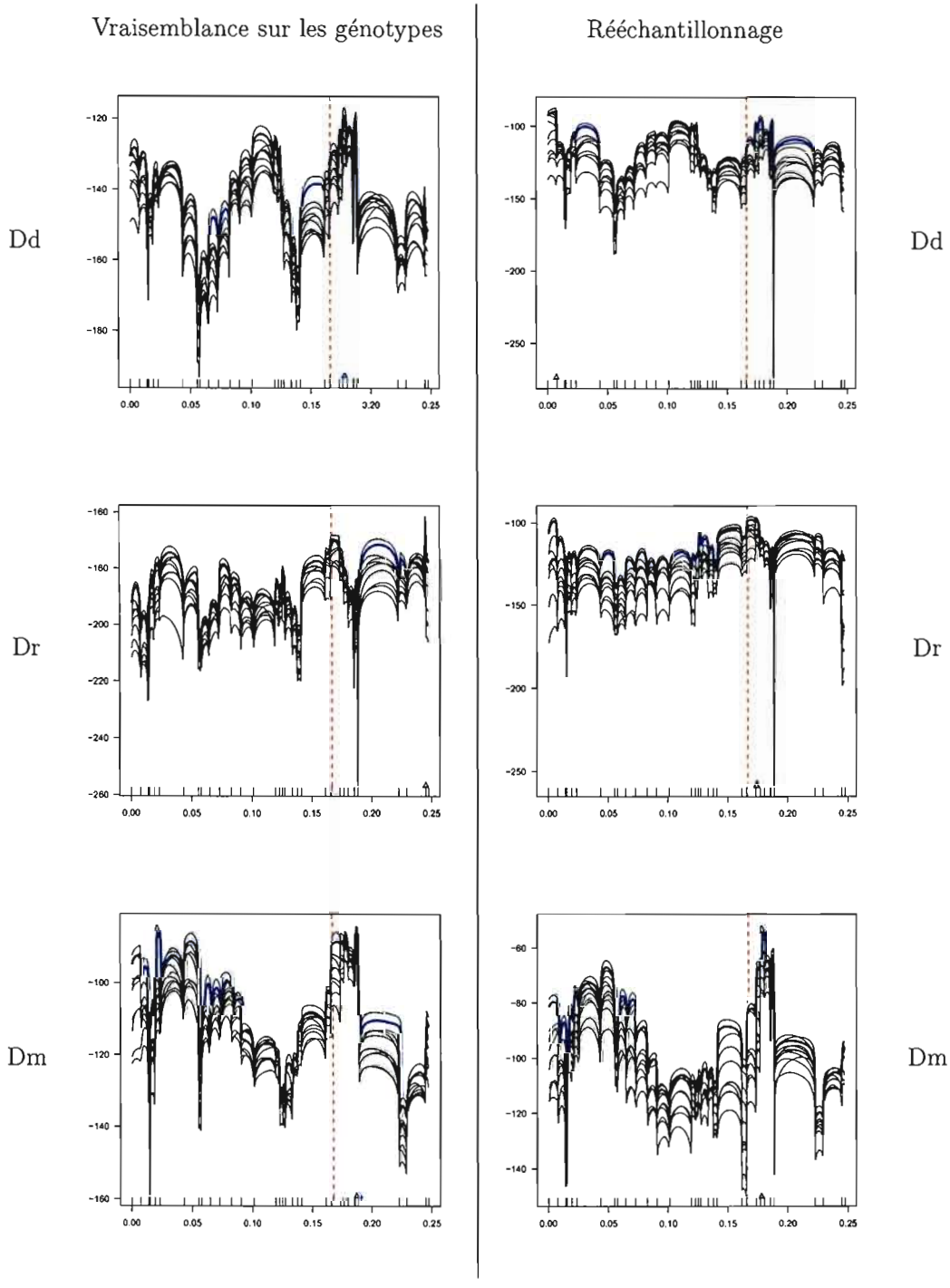

Figure 6.14 MapArg sur les données D avec des fenêtres de 4 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est supposé connu. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

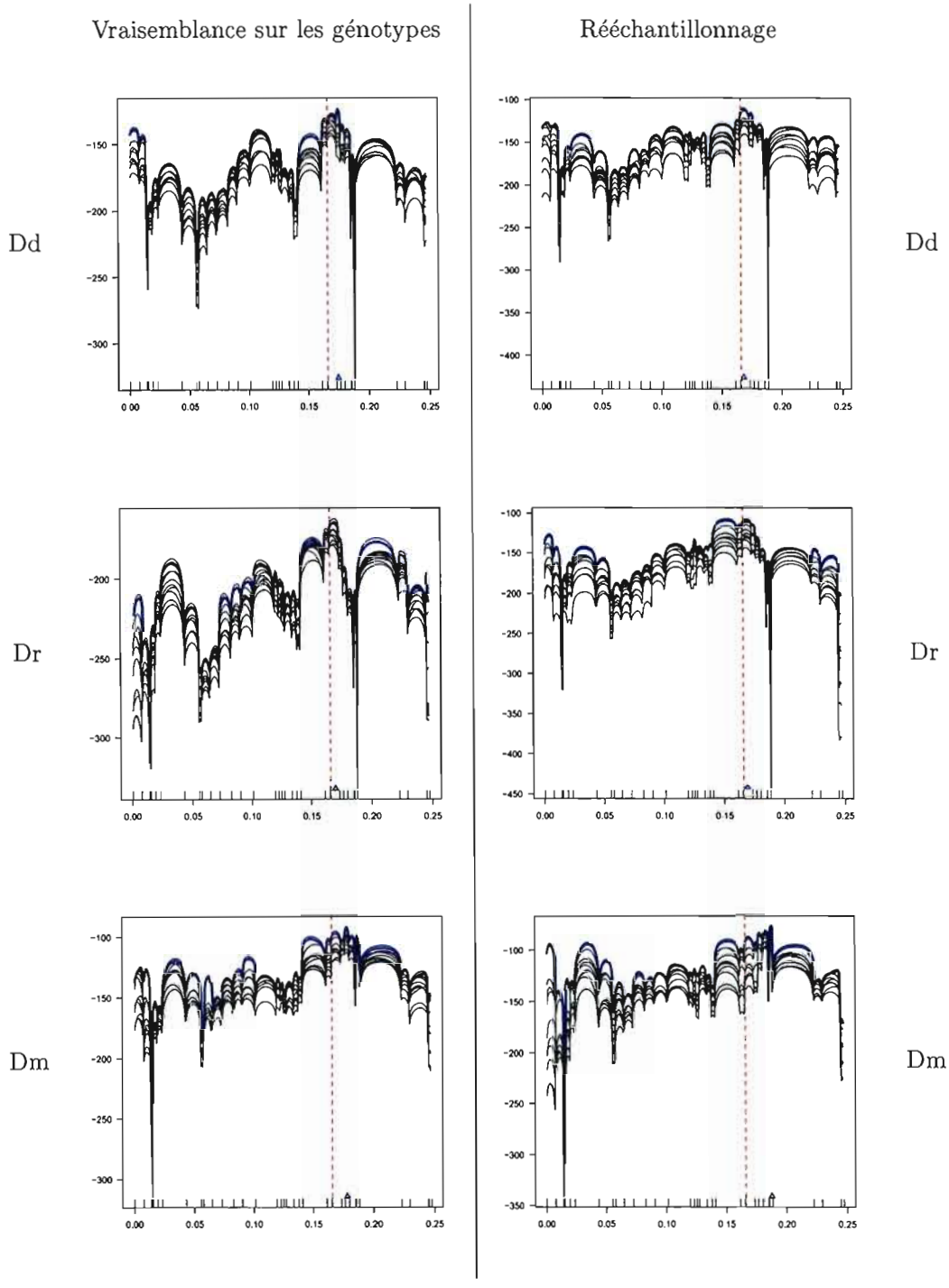

Figure 6.15 MapArg sur les données D avec des fenêtres de 6 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est supposé connu. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

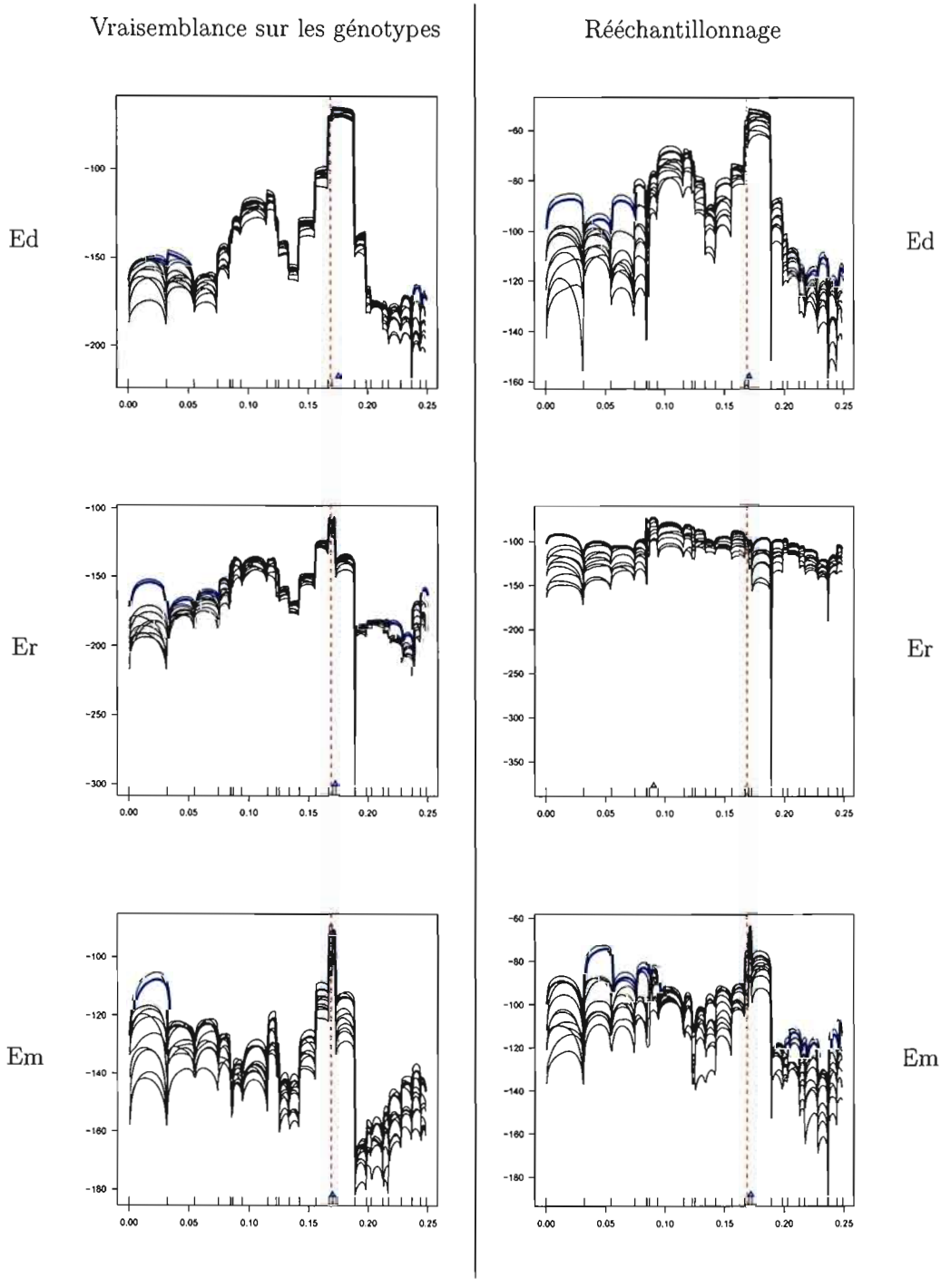

Figure 6.16 MapArg sur les données E avec des fenêtres de 4 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est supposé connu. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

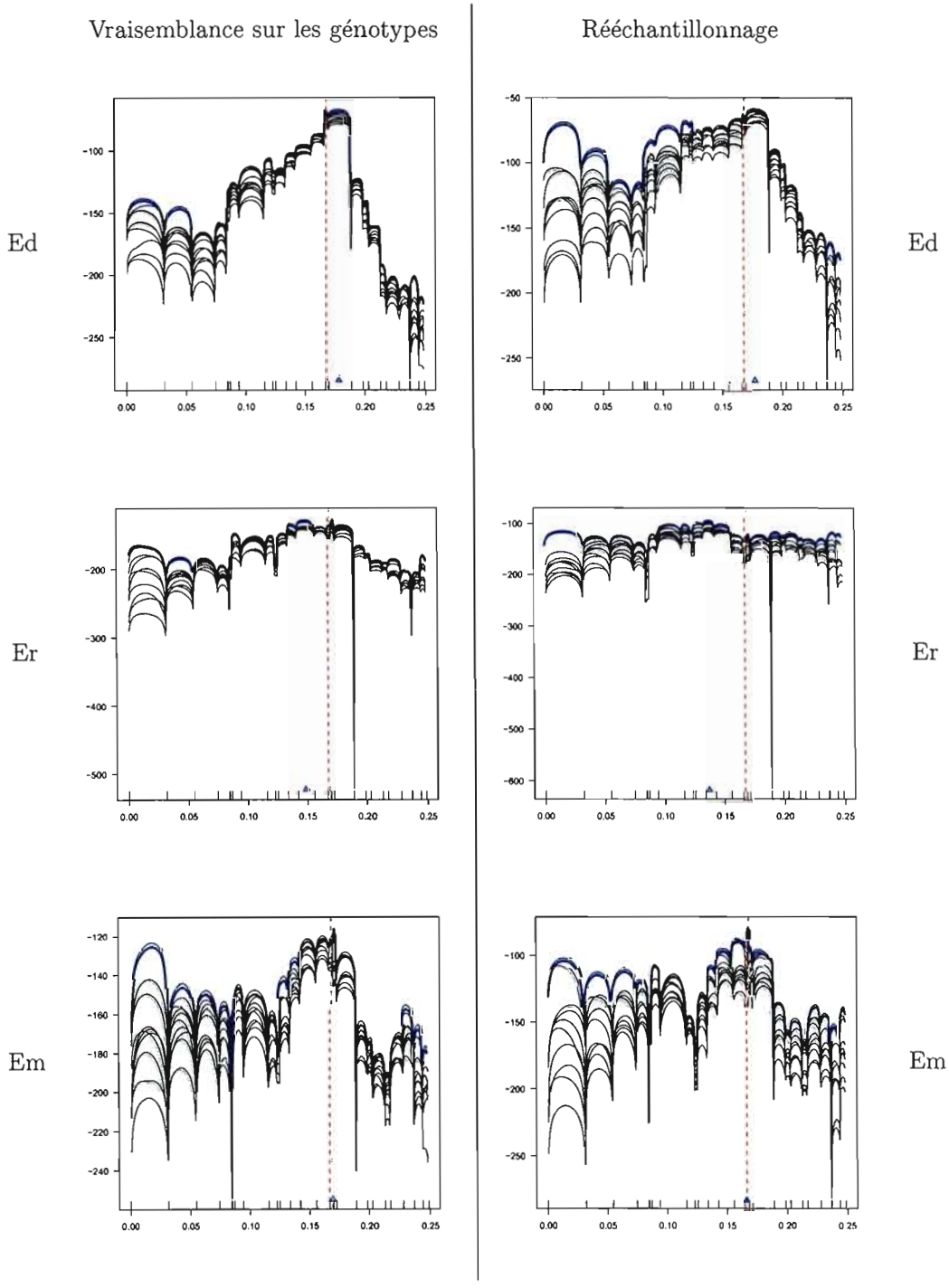

Figure 6.17 MapArg sur les données E avec des fenêtres de 6 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est supposé connu. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

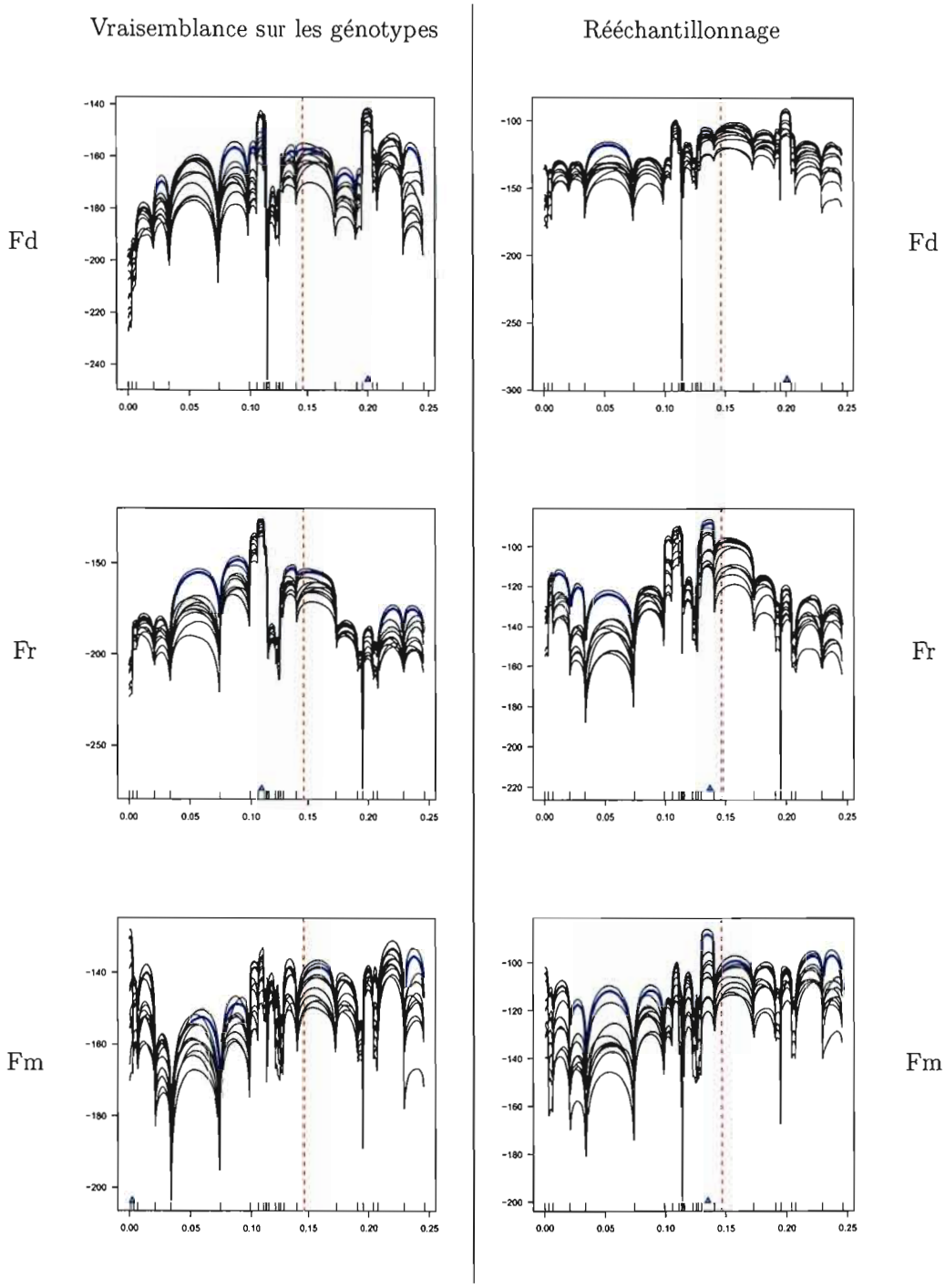

Figure 6.18 MapArg sur les données F avec des fenêtres de 4 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est supposé connu. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

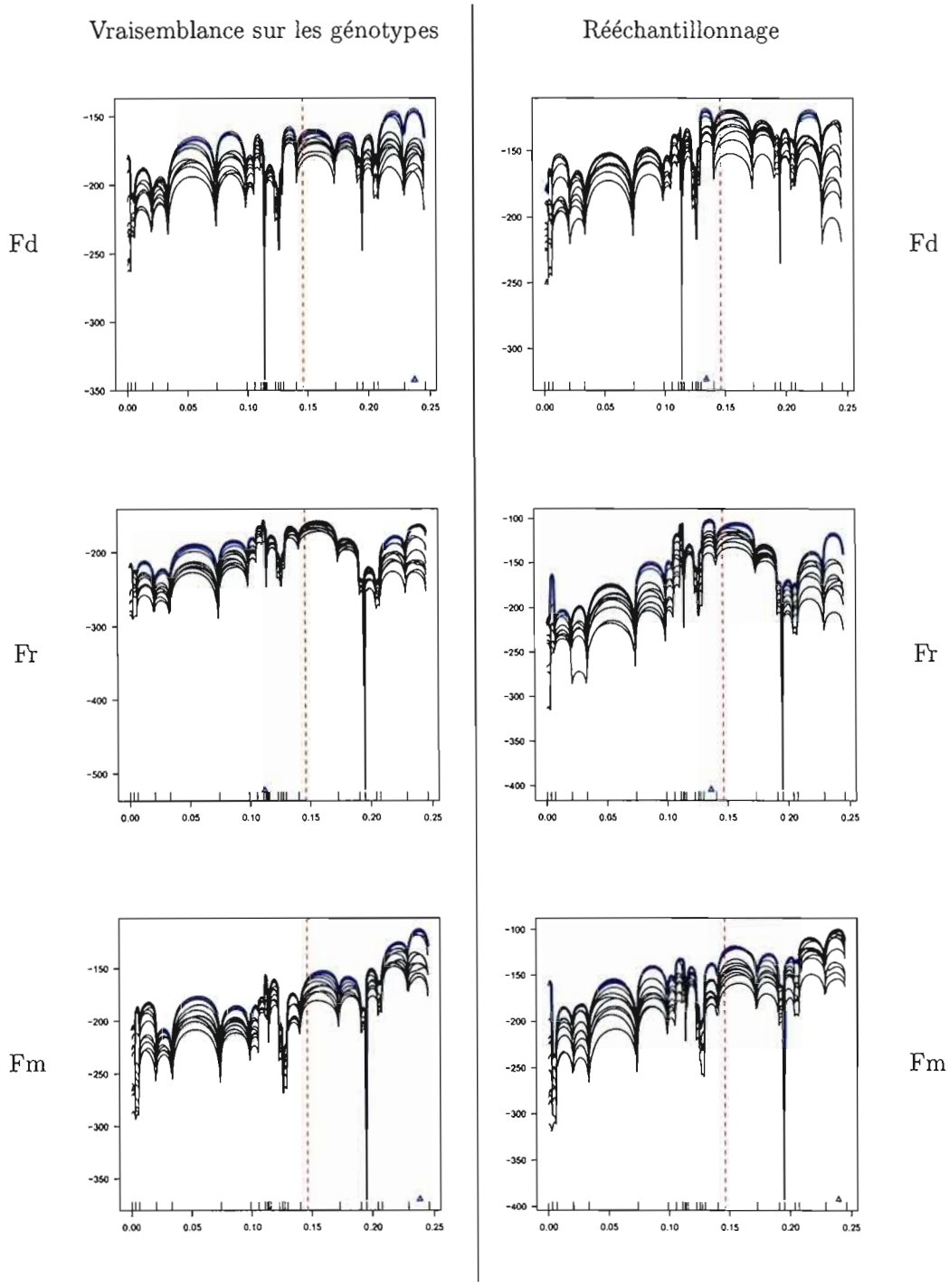

Figure 6.19 MapArg sur les données F avec des fenêtres de 6 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est supposé connu. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

Les résultats suggèrent que l'intégration de la réalité diploïde par l'algorithme EM est nettement préférable à une intégration arbitraire lorsque le modèle est connu. Ainsi, pour les données B, D et E, on obtient des résultats presque aussi précis que lorsque les séquences haploïdes sont disponibles. Pour les données F, par contre, les résultats sont plus variables. Ceci concorde avec les graphiques de distances de la figure 6.9. En effet, on y note une augmentation importante de la distance entre V_0 et V_1 à proximité du gène causal pour les données B, D et E, ce qui dénote un déséquilibre de liaison plus important. On peut donc supposer que ces données contiennent plus d'information sur la position de la mutation. Ainsi, malgré le bruit introduit par l'intégration de la réalité diploïde, il demeure suffisamment d'information pour détecter la position du gène.

Nous avons présenté deux méthodes d'intégration de la réalité diploïde. Un des objectif de cette section est de comparer les performances de celles-ci. Il semblerait que le calcul de la vraisemblance sur les génotypes et phénotypes soit préférable au rééchantillonnage. Précisons cependant que les résultats obtenus par cette dernière méthode dépendent du nombre de séquences porteuses et non porteuses choisies. Enfin, notons que la précision est légèrement augmentée lorsque des fenêtres de 6 marqueurs, plutôt que 4, sont utilisées. L'algorithme EM performe moins bien lorsque la taille des fenêtre augmente, tandis que la méthode MapArg perd de la précision lorsque celle-ci diminue. Les fenêtres de 6 marqueurs nous semblent donc optimales. D'autres essais, non illustrés ici, nous ont permis de constater que les fenêtres de 8 marqueurs n'étaient pas appropriées, en raison des erreurs induites par l'algorithme EM.

6.3 Modèle de pénétrance inconnu

6.3.1 Estimation des modèles

Les tests que nous avons effectués précédemment supposaient la connaissance du modèle de pénétrance. En pratique, ce dernier n'est généralement pas connu. Il nous faut donc l'évaluer. Nous avons construit au chapitre 4 un algorithme permettant d'estimer le modèle de pénétrance à partir des génotypes et phénotypes. Rappelons que ce dernier

est basé sur un algorithme EM et évalue la vraisemblance d'un ensemble de modèles compatibles avec la fréquence observée du caractère d'intérêt. Cet ensemble de modèle a été choisi de manière à ce que la probabilité d'observer le caractère augmente avec le nombre de mutations portées, c'est-à-dire que $f_0 \leq f_1 \leq f_2$. De plus, nous avons appliqué des contraintes sur les paramètres :

$$\begin{aligned} 0,001 \leq \quad & f_0 \quad \leq 0,200; \\ 0,001 \leq \quad & f_1 \quad \leq 0,999; \\ 0,600 \leq \quad & f_2 \quad \leq 0,999; \\ 0,010 \leq \quad & p \quad \leq 0,200. \end{aligned}$$

Ce choix de contraintes reflète une connaissance préalable minime sur la maladie. Par exemple, l'intervalle choisi pour la proportion p de séquences porteuses dans la population correspond à l'hypothèse que le gène est relativement rare.

L'algorithme d'estimation de modèle ne peut être appliqué que sur une fenêtre restreinte de marqueurs à la fois. Nous avons donc découpé les séquences en fenêtres et avons évalué la vraisemblance de chaque modèle sur chacune des fenêtres. Le modèle choisi correspond au maximum de toutes ces vraisemblances estimées. Nous avons répété l'opération pour des fenêtres de 5, 6 et 7 marqueurs. Les tableaux 6.2 à 6.5 résument les résultats obtenus pour les paramètres des différents modèles.

De manière générale, le taux de phénocopies est sous-évalué dans les modèles estimés. Ainsi, f_0 atteint presque toujours la valeur minimale admise. De même, la fréquence de la mutation est généralement surestimée. On note aussi que l'erreur est parfois importante sur les paramètres. Une meilleure évaluation préalable de la fréquence de la mutation ou du taux de phénocopies améliorerait la qualité de l'estimation pour les autres paramètres. Notons que, malgré les erreur sur les paramètres, on distingue assez facilement les modèles dominants et récessifs.

Tableau 6.2 Modèles de pénétrance estimés, données B

Bd

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,010	0,900	0,950	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,439	0,999	0,196
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,439	0,999	0,196
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,950	0,960	0,097

Br

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,010	0,010	0,950	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,001	0,600	0,163
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,001	0,600	0,163
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,001	0,600	0,163

Bm

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,050	0,100	0,800	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,013	0,025	0,986	0,164
Modèle estimé : fenêtre de 6 marqueurs	0,013	0,025	0,845	0,177
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,050	0,703	0,195

Tableau 6.3 Modèles de pénétrance estimés, données D

Dd

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,010	0,900	0,950	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,926	0,935	0,101
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,634	0,638	0,152
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,634	0,638	0,152

Dr

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,010	0,010	0,950	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,001	0,613	0,161
Modèle estimé : fenêtre de 6 marqueurs	0,013	0,025	0,986	0,048
Modèle estimé : fenêtre de 7 marqueurs	0,013	0,025	0,948	0,048

Dm

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,050	0,100	0,800	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,074	0,999	0,198
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,074	0,973	0,199
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,074	0,973	0,199

Tableau 6.4 Modèles de pénétrance estimés, données E

Ed

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,010	0,900	0,950	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,439	0,999	0,199
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,439	0,999	0,199
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,585	0,600	0,167

Er

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,010	0,010	0,950	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,001	0,600	0,167
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,001	0,600	0,167
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,001	0,896	0,137

Em

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,050	0,100	0,800	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,063	0,074	0,999	0,012
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,074	0,999	0,199
Modèle estimé : fenêtre de 7 marqueurs	0,063	0,074	0,999	0,012

Tableau 6.5 Modèles de pénétrance estimés, données F

Fd

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,010	0,900	0,950	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,101	0,926	0,935	0,196
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,561	0,960	0,197
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,634	0,639	0,152

Fr

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,010	0,010	0,950	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,001	0,613	0,161
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,001	0,600	0,167
Modèle estimé : fenêtre de 7 marqueurs	0,013	0,025	0,948	0,048

Fm

Paramètres	f_0	f_1	f_2	p
Modèle véritable	0,050	0,100	0,800	0,100
Modèle estimé : fenêtre de 5 marqueurs	0,001	0,195	0,703	0,154
Modèle estimé : fenêtre de 6 marqueurs	0,001	0,147	0,639	0,189
Modèle estimé : fenêtre de 7 marqueurs	0,001	0,098	0,986	0,193

6.3.2 Impact de l'estimation préalable du modèle

Nous avons estimé dans la section précédente les paramètres du modèle de pénétrance associé à chaque échantillon, avec différentes longueur de fenêtres. Afin d'évaluer l'impact de l'estimation préalable et ponctuelle du modèle, nous avons appliqué la méthode MapArg en utilisant les paramètres estimés. Nous avons choisi pour ce faire des fenêtres de 6 marqueurs. Les figures 6.20 à 6.23 illustrent les résultats.

On note une grande variabilité dans les résultats. Évidemment, la connaissance exacte du modèle apportait une information supplémentaire et, par conséquent, des estimations plus exactes de la position de la mutation. Dans la mesure où les paramètres estimés présentent des erreurs parfois importantes, il n'est donc pas surprenant que certaines courbes de vraisemblance soient inexactes. Cependant, il ne semble pas y avoir de corrélation directe entre la qualité de l'estimation du modèle de pénétrance et celle de la position de la mutation. Ceci peut s'expliquer par le fait que l'algorithme EM ne dépend pas directement des paramètres f_0, f_1 et f_2, mais plutôt des probabilités conditionnelles aux phénotypes décrites à la page 66.

Les tests que nous avons effectués ne nous permettent pas de favoriser une méthode d'intégration plutôt qu'une autre lorsque le modèle est estimé. Le calcul de la vraisemblance sur les génotypes est plus directement dépendant du modèle. Par contre, puisque celui-ci intègre l'estimation des haplotypes par l'ajout d'étapes à la méthode d'échantillonnage pondéré, il est possible que l'impact du modèle soit atténué. En effet, le même modèle est utilisé pour le calcul de la vraisemblance et pour la distribution proposée. Ainsi, la vraisemblance est aussi pondérée par le modèle, ce qui minimise peut-être l'impact de ce dernier.

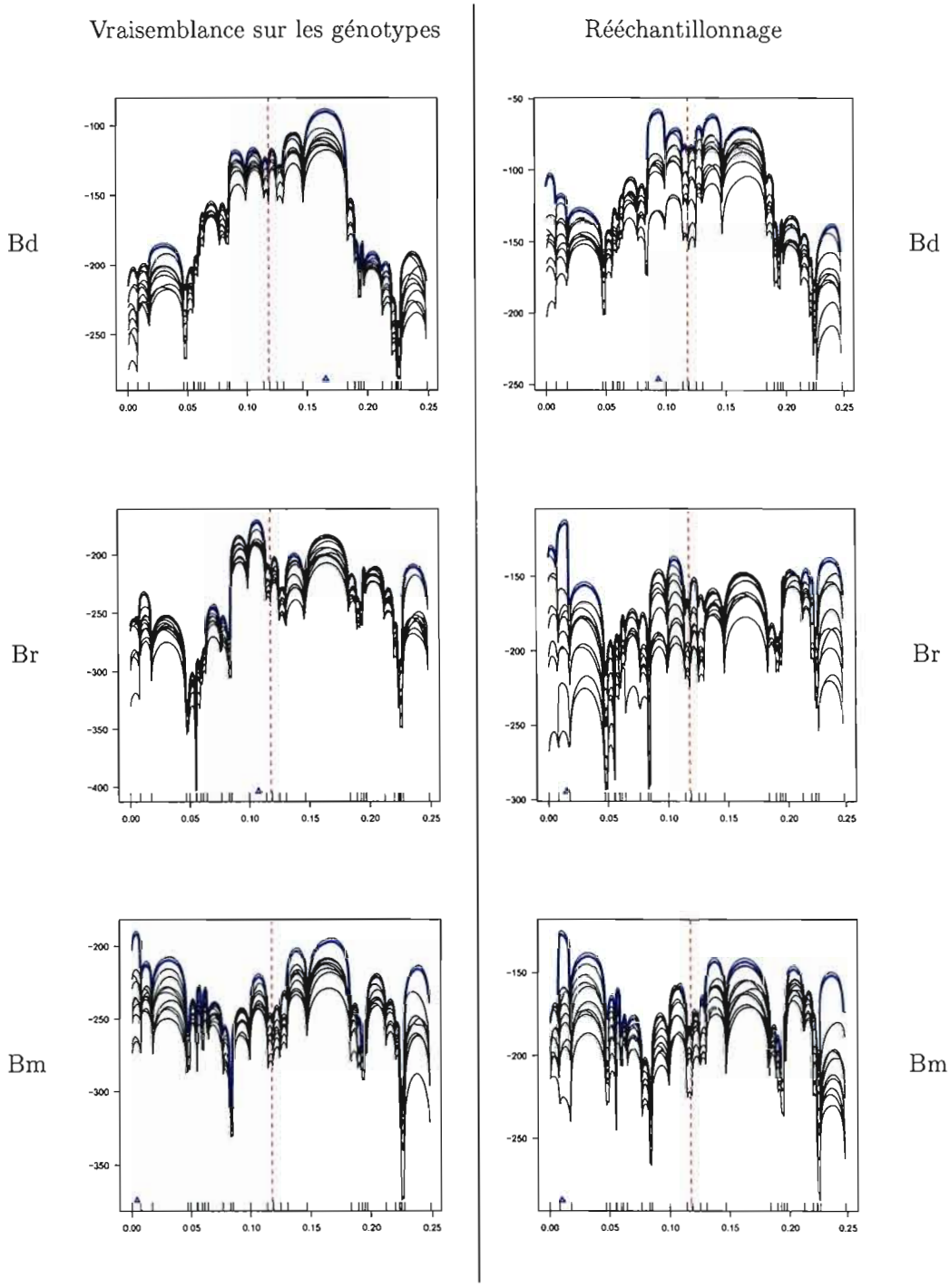

Figure 6.20 MapArg sur les données B avec des fenêtres de 6 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est estimé. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

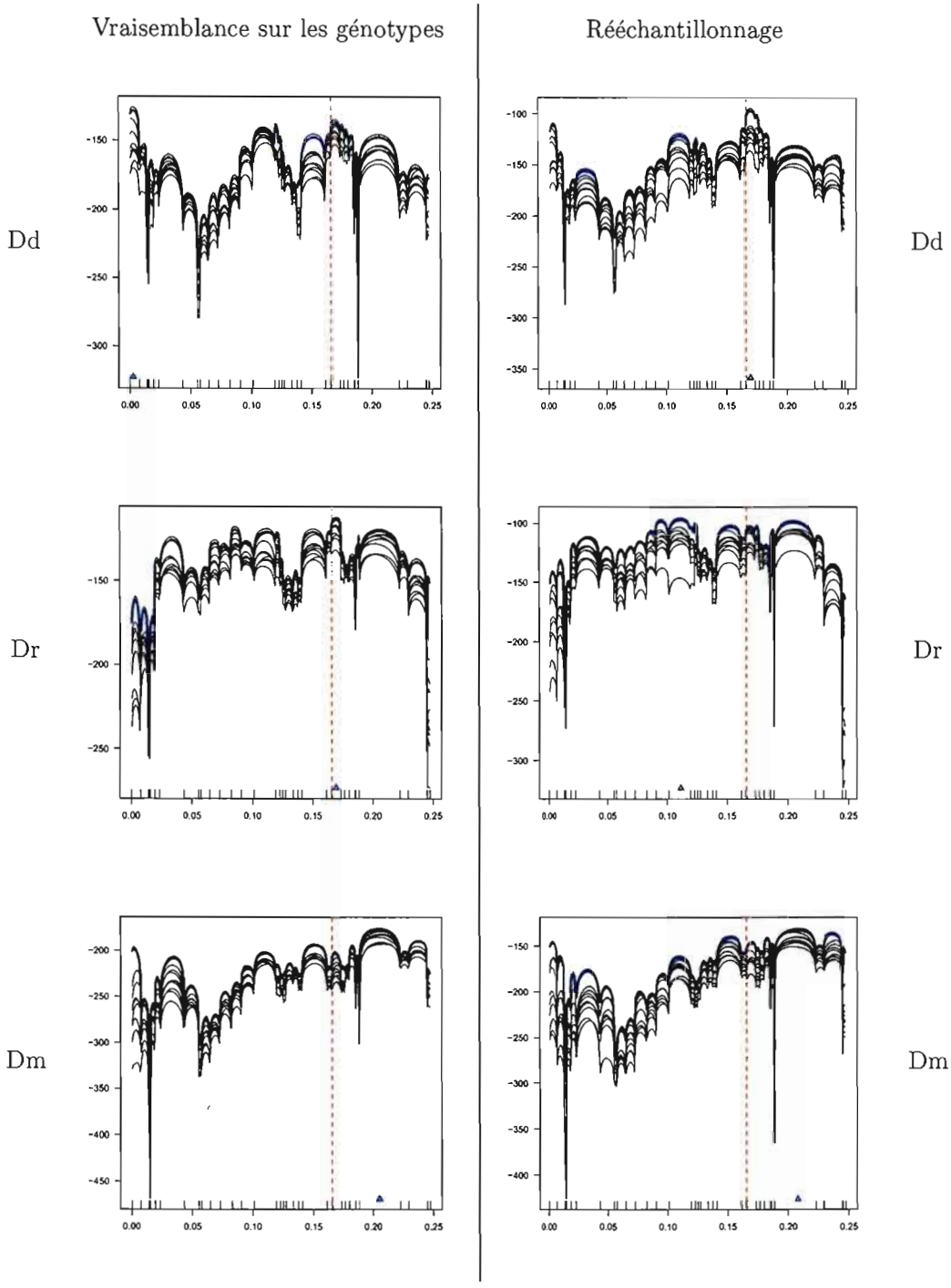

Figure 6.21 MapArg sur les données D avec des fenêtres de 6 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est estimé. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

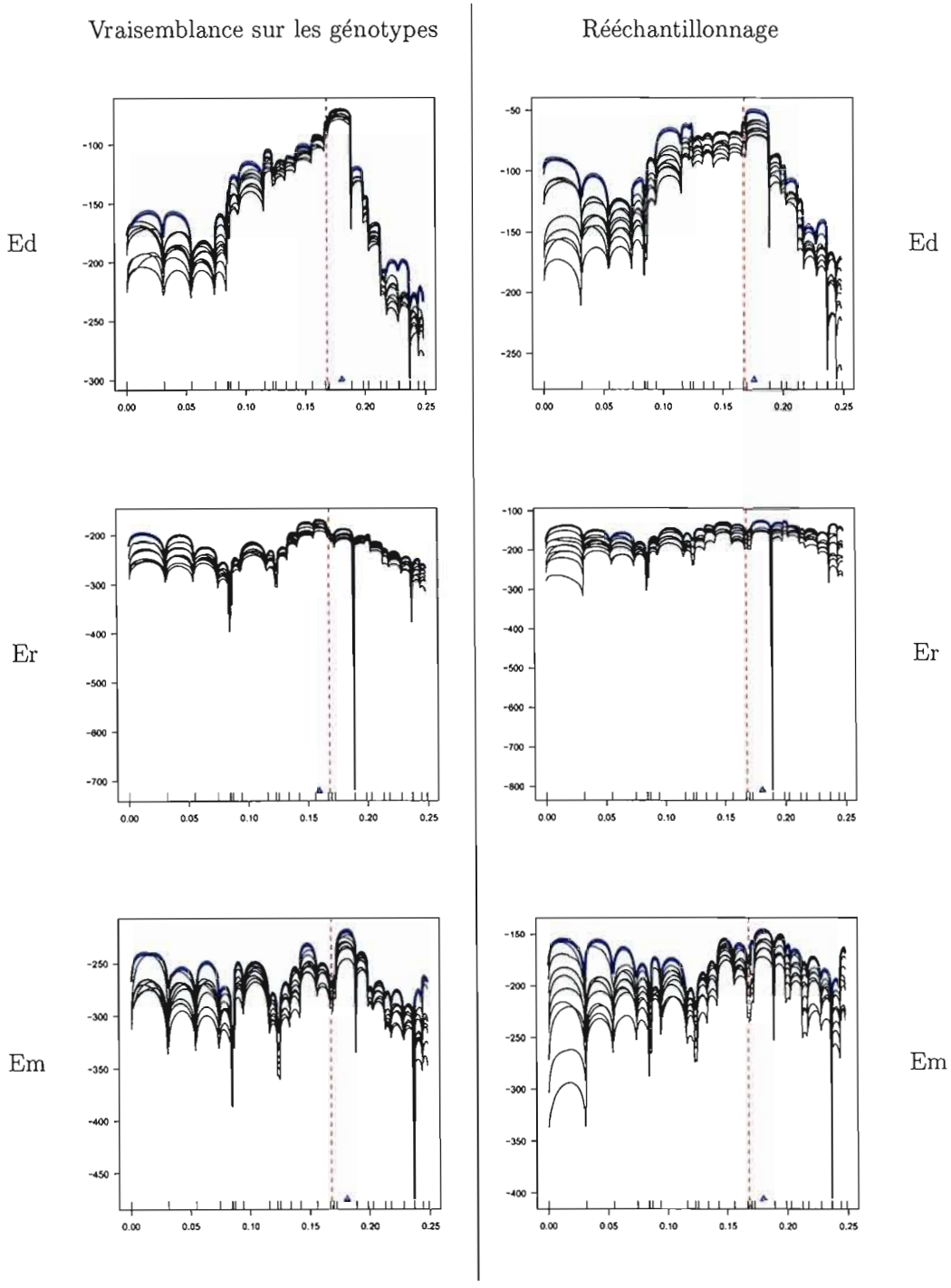

Figure 6.22 MapArg sur les données E avec des fenêtres de 6 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est estimé. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

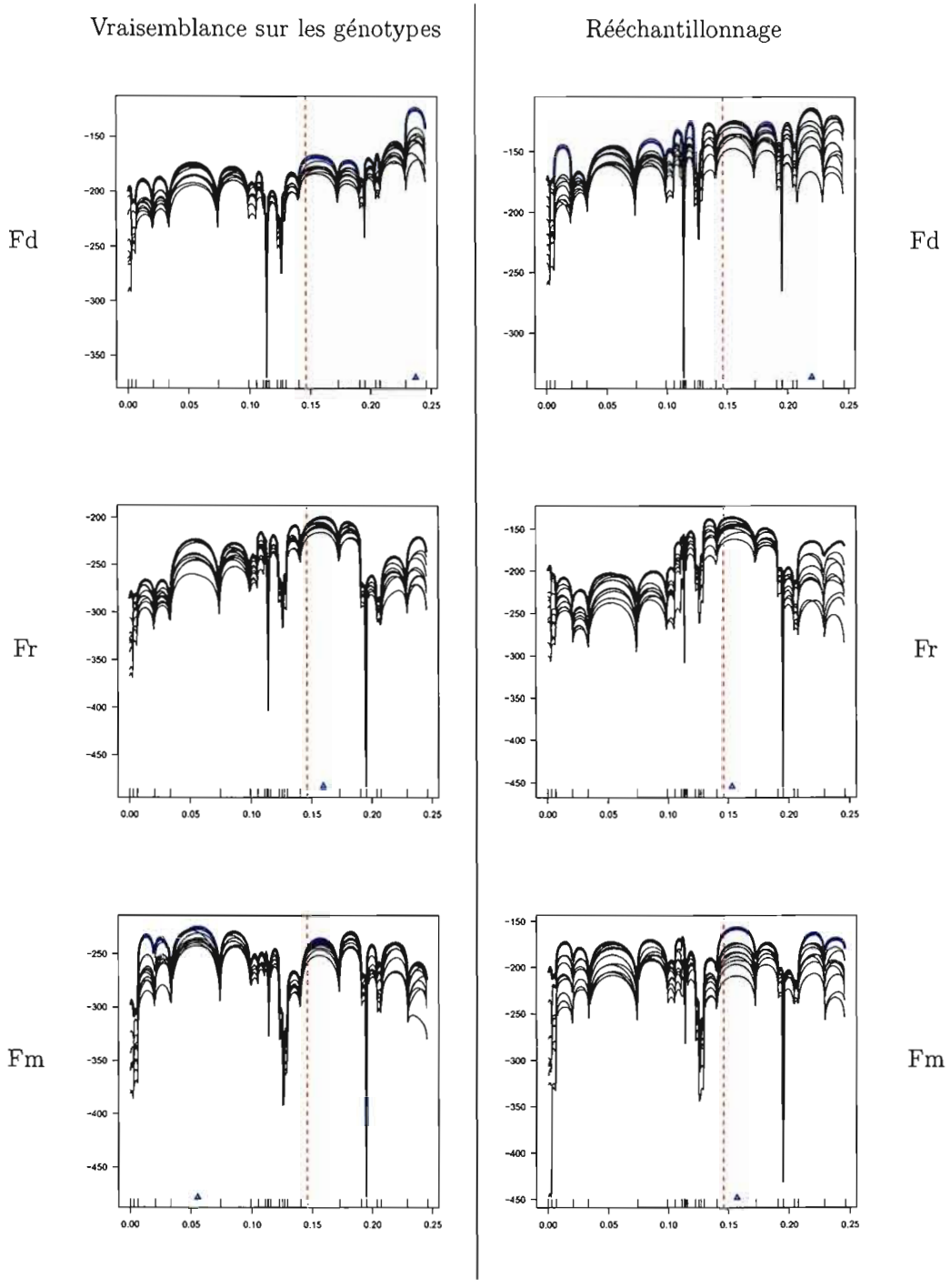

Figure 6.23 MapArg sur les données F avec des fenêtres de 6 marqueurs. À gauche, les résultats obtenus par le calcul de la vraisemblance sur les génotypes. À droite, les résultats obtenus par rééchantillonnage. De haut en bas : un modèle dominant, récessif et mixte avec respectivement 1%, 1% et 5% de phénocopies. Le modèle de pénétrance est estimé. La ligne verticale est la position réelle de la mutation. Le triangle est la position estimée.

Ce chapitre nous a permis d'évaluer les performances de l'algorithme EM conditionnel aux phénotypes, ainsi que l'intégration de la réalité diploïde à la méthode MapArg. Évidemment, les résultats présentés ne sont que des exemples obtenus sur des données simulées. De plus, l'évaluation s'est faite de manière subjective, à l'aide de graphiques. Toutefois, les exemples proposés suggèrent que l'algorithme EM performe plutôt bien lorsque la taille des fenêtres est raisonnable (de quatre à six marqueurs). De même, les deux techniques d'intégration de la réalité diploïde semblent adéquates lorsque ces tailles de fenêtres sont suffisantes pour la méthode MapArg. Dans tous les cas, cependant, une bonne connaissance du modèle de pénétrance est nécessaire. Bien que l'algorithme d'estimation des modèles que nous avons développé ait relativement bien performé, il semble hasardeux d'utiliser une estimation ponctuelle dans la méthode MapArg. Ceci n'est pas surprenant, dans la mesure où l'incertitude sur les paramètres n'est alors pas prise en compte. Il serait certainement plus adéquat de considérer une grille de modèles, ce qui constitue une nouvelle piste de travail pour des améliorations futures.

CONCLUSION

Nous avions pour objectif de généraliser une méthode de cartographie génétique fine afin d'intégrer la réalité diploïde, de manière à permettre son application directe à des échantillons de génotypes et phénotypes humains, plutôt qu'à des haplotypes uniquement. Par ailleurs, nous voulions permettre son application à une vaste gamme de modèles de pénétrance. Il s'agit d'un pas important, puisque la méthode originale était restreinte à des conditions d'application particulières et contraignantes. La méthode généralisée devrait quant à elle pouvoir être utilisée pour la cartographie génétique d'un plus vaste ensemble de caractères héréditaires, en plus d'être directement applicable aux échantillons de génotypes obtenus en laboratoire.

La méthode originale vise à estimer la position d'une mutation causale d'un caractère héréditaire par la simulation de graphes de recombinaison ancestraux. Celle-ci supposait toutefois un échantillon de séquences haploïdes pour lesquelles le statut au gène causal est connu, ce qui est rarement possible. Afin de généraliser la méthode, nous avons d'abord mis en évidence le modèle d'échantillonnage pondéré de celle-ci. Nous avons ensuite exploré différentes solutions au problème de l'estimation des haplotypes. Par la suite, nous avons développé un nouvel algorithme EM prenant en considération les phénotypes et le modèle de pénétrance. Cet algorithme permet d'estimer les distributions d'haplotypes parmi les séquences porteuses et non porteuses de la population. Il devient alors possible d'estimer les séquences haploïdes correspondant au génotype d'un individu, ainsi que le statut au gène causal. Nous avons finalement intégré la réalité diploïde par l'ajout d'étapes au modèle d'échantillonnage pondéré de la méthode.

Des tests par simulation nous ont permis de constater dans un premier temps que l'algorithme EM que nous avons développé estimait plutôt bien les fréquences d'haplotypes. L'application de l'algorithme est toutefois limitée à une taille de fenêtre restreinte,

puisque le nombre de paramètres à estimer croît de façon exponentielle avec le nombre de marqueurs impliqués. Nos tests laissent supposer que la taille optimale se situe entre quatre et six marqueurs. De ce fait, l'application de la méthode MapArg généralisée est aussi limitée à l'utilisation de fenêtres de cette taille. L'utilisation de la vraisemblance composite permet toutefois de couvrir l'ensemble de la séquence, par petites fenêtres consécutives. La performance de la méthode MapArg intégrant la réalité diploïde par l'algorithme EM a aussi été évaluée par simulation. Les résultats sont concluants lorsque le modèle de pénétrance est supposé connu. Nous avons comparé des tailles de fenêtres de quatre et six marqueurs. Il semblerait que des tailles de six marqueurs soient préférables. Nous avons aussi effectué quelques tests en utilisant des modèles de pénétrance estimés. Dans ce cas, les résultats obtenus varient beaucoup. Évidemment, la performance est meilleure lorsque le modèle est connu.

Nous avons choisi d'utiliser un algorithme EM afin d'estimer les haplotypes à partir des génotypes des individus. Il serait toutefois intéressant de tenter d'intégrer la réalité diploïde en appliquant une autre méthode, telle une variante de Phase. Il serait aussi souhaitable de travailler davantage la problématique des modèles de pénétrance inconnus, que nous n'avons que survolée. Par exemple, il serait possible de considérer un grand nombre de modèles, en supposant une distribution *a priori* sur ceux-ci. Enfin, davantage de tests devraient être effectués. Il serait entre autre essentiel d'évaluer les performances de la méthode généralisée à des échantillons d'ADN humain, plutôt qu'à des données simulées selon le processus de coalescence.

APPENDICE A

DÉTAILS DES PROBABILITÉS D'ÉVÉNEMENTS DANS MAPARG

Cette section est un complément d'information concernant le calcul des probabilités des divers événements dans MapArg (voir page 24).

Considérons un événement de coalescence C_{ij}^{k}. Dans cette situation, il y avait à l'étape $H_{\tau+1}$ une séquence de type k supplémentaire, mais une séquence de type i et une de type j en moins. Si $i = k$, il y avait uniquement une séquence de type j en moins. Notons $\delta_{ik} = 1$ si $i = k$, 0 sinon. Le nombre de séquences de type k dans le passé est alors donné par $n_k + 1 - \delta_{ik} - \delta_{jk}$. On a alors que

$$Q_{r_\tau}(H_\tau \mid H_{\tau+1}) = P_\tau(C)\frac{(n_k + 1 - \delta_{ik} - \delta_{jk})}{(n-1)}$$

où $H_{\tau+1} = H_\tau + C_{ij}^{k}$. Remarquons que, si $i = j = k$, on se retrouve dans le cas d'une coalescence identique.

Dans le cas d'une mutation, le nombre de séquences demeure identique. Ainsi, pour une mutation $M_i^j(m)$, il y avait à l'étape $H_{\tau+1}$ une séquence de type j en plus et une de type i en moins. La probabilité de choisir une séquence de type j pour la mutation est alors de $(n_j + 1)/n$. La probabilité que la mutation soit survenue au marqueur m est quant à elle de $\theta_m/(\alpha\theta)$. Ceci nous donne

$$Q_{r_\tau}(H_\tau \mid H_{\tau+1}) = P_\tau(M)\frac{\theta_m}{\alpha\theta}\frac{(n_j+1)}{n},$$

où $H_{\tau+1} = H_\tau + M_i^j(m)$.

Lors d'une recombinaison, le nombre total de séquences est augmenté de un. Ainsi, pour une recombinaison R_i^{jk}, il y avait dans le passé une séquence de type i en moins, ainsi qu'une de type j et une de type k en plus. Notons qu'une recombinaison implique toujours des séquences différentes et que l'ordre des séquences parentales est important. En effet, l'une transmet le matériel à droite du point de recombinaison, tandis que l'autre transmet le matériel à gauche de ce même point. La probabilité qu'une séquence de type j et une séquence de type k soient impliquées dans une recombinaison est alors de $(n_j + 1)(n_k + 1)/(n(n + 1))$. La probabilité que la recombinaison survienne dans l'intervalle m est quant à elle donnée par $\rho_m(r_\tau)/[\rho\beta(r_\tau)]$. On retrouve alors la probabilité

$$Q_{r_\tau}(H_\tau \mid H_{\tau+1}) = P_\tau(R)\frac{\rho_m(r_\tau)}{\rho\beta(r_\tau)}\frac{(n_j+1)(n_k+1)}{n(n+1)},$$

où $H_{\tau+1} = H_\tau + R_i^{jk}(m)$.

APPENDICE B

NOTATIONS : MÉTHODE MAPARG

L	Nombre total de marqueurs
x_i	Position du marqueur i
r_i	Distance génétique entre les marqueurs i et $i+1$
r_T	Position de la mutation causale
r_{T_0}	Valeur conductrice pour la position de la mutation causale
H_i	Ensemble des séquences haploïdes à l'étape i de la chaîne
H_0	Ensemble des séquences haploïdes observées
H_{τ^*}	Ancêtre commun à toutes les séquences observées ($MRCA$)
$P(H_i \mid H_{i-1})$	Probabilité conditionnelle à rebours dans le temps
$Q(H_i \mid H_{i+1})$	Probabilité conditionnelle dans le sens chronologique
C_i	Coalescence identique de deux séquences de type i
C_{ij}^k	Coalescence non identique de séquences de type i et j vers une séquence parentale de type k
$M_i^j(m)$	Mutation d'une séquence de type i vers une séquence parentale de type j, au marqueur m
$R_i^{jk}(m)$	Recombinaison d'une séquence de type i vers des séquences parentales j et k, dans l'intervalle m
μ_m	Taux de mutation au marqueur m, par génération
θ_m	Taux de mutation au marqueur m, à l'échelle de coalescence
r_m	Taux de recombinaison dans l'intervalle m, par génération
ρ_m	Taux de recombinaison dans l'intervalle m, à l'échelle de coalescence

n_i	Fréquence des séquences de type i, à une étape donnée
n	Nombre total de séquences à une étape donnée
$A^{(i)}$	Ensemble des marqueurs ancestraux d'une séquence i, à une étape donnée
$B^{(i)}$	Ensemble des intervalles situés sur du materiel ancestral d'une séquence i, à une étape donnée
α	Facteur de correction du taux global de mutation, en fonction du matériel ancestral
β	Facteur de correction du taux global de recombinaison, en fonction du matériel ancestral
A_j	Marqueurs et intervalle de la fenêtre j
H_0^*	Séquences observées, sans la mutation causale

APPENDICE C

NOTATIONS : ALGORITHME EM

G	Échantillon ordonné de génotypes
Φ	Échantillon ordonné de phénotypes
D	Diplotypes parentaux (inconnus) correspondant à l'échantillon
g	Génotype d'un individu
ϕ	Phénotype d'un individu ($\phi = 1$ pour cas, $\phi = 0$ pour témoin)
$V(h)$	Distribution des haplotypes dans la population haploïde
$V_0(h)$	Distribution des haplotypes non mutants
$V_1(h)$	Distribution des haplotypes mutants
h^0	Haplotype de type h non mutant
h^1	Haplotype de type h mutant
m_h	Fréquence de séquences de type h dans D
$m_h^{(k)}$	Fréquence moyenne de séquences de type h, après l'itération k
n_g	Fréquence du génotype g dans l'échantillon
n_ϕ	Fréquence du phénotype ϕ dans l'échantillon
$n_{g,\phi}$	Fréquence du génotype g et du phénotype ϕ dans l'échantillon
F	Modèle de pénétrance ; $F = (f_0, f_1, f_2)$
f_i	Probabilité qu'un individu portant i copies du gène causal soit atteint
p	Proportion de séquences haploïdes porteuses du gène causal
f	Fréquence du trait dans la population diploïde

T	Statut d'un individu au gène causal
	$T = 00$, individu non porteur ;
	$T = 01$, haplotype paternel porteur ;
	$T = 10$, haplotype maternel porteur ;
	$T = 11$, individu doublement porteur.
$[h_i, h_j]$	Diplotype parental : h_i est l'haplotype maternel, h_j est le paternel
$[h_i, h_j] \in g$	Diplotype parental compatible avec le génotype g
(h_i, h_j)	Diplotype au sens usuel : paire d'haplotypes

APPENDICE D

NOTATIONS : INTÉGRATION DE LA RÉALITÉ DIPLOÏDE

H_{-2}	Ensemble ordonné des génotypes et phénotypes observés
H_{-1}	Ensemble ordonné des diplotypes des individus
H_0	Ensemble non ordonné des séquences haploïdes
n	Nombre d'individus diploïdes dans l'échantillon
m_s	Fréquence des séquences de type s dans H_0
ω	Proportion fixée de cas dans l'échantillon
α_i	Proportion attendue d'individus cas parmi ceux portant $i \in \{0,1,2\}$ copies de la mutation causale, dans un échantillon à proportion fixée de cas
q_i	Proportion attendue d'individus portant $i \in \{0,1,2\}$ copies de la mutation causale
q	Fréquence attendue de la mutation causale dans l'échantillon
N_i	Nombre d'individus portant $i \in \{0,1,2\}$ copies de la mutation causale dans H_{-1}
M	Nombre de séquences porteuses dans H_0

RÉFÉRENCES

Almgren, P., P.-O. Bendahl, H. Bengtsson, O. Hösser, et R. Perfekt. 2003. *Statistics in genetics (lecture notes)*. Centre for Mathematical Sciences, Lund University.

Campbell, N. A. et R. Mathieu. 1995. *Biologie*. Éditions du renouveau pédagogique.

Clark, A. G. 1990. « Inference of haplotypes from pcr-amplified samples of diploid populations », *Molecular Biology Evolution*, vol. 7, no. 2, p. 111–122.

Dempster, A., N. Laird, et D. Rubin. 1977. « Maximum likelihood from incomplete data via the em algorithm », *Journal of the royal statistical society*, vol. 39, p. 1–38.

Excoffier, L. et M. Slatkin. 1995. « Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population », *Molecular Biology Evolution*, vol. 12, no. 5, p. 921–927.

Griffiths, R. C. et P. Marjoram. 1996. « Ancestral inference from samples of dna sequences with recombination », *Journal of Computational Biology*, vol. 3, no. 4, p. 479–502.

———. 1997. *An ancestral recombination graph*. Coll. Donnelly, P. et S. Tavaré, éditeurs, Coll. « *Progress in Population Genetics and Human Evolution* ». Springer-Verlag.

Hein, J., M. H. Schierup, et C. Wiuf. 2005. *Gene Genealogies, variation and evolution, a primer un coalescent theory*. Oxford University Press.

Huang, Y.-T., K.-M. Chao, et C. Ting. 2005. « An approximation algorithm for haplotype inference by maximum parsimony », *Journal of Computational Biology*, vol. 12, no. 10, p. 1261–1274.

Hudson, R. 2002. « Generating samples under a wright-fisher neutral model of genetic variation », *Bioinformatics*, vol. 18, p. 337–338.

Ito, T., E. Inoue, et N. Kamatani. 2004. « Association test algorithm between a qualitative phenotype and a haplotype or haplotype set using simultaneous estimation of haplotype frequencies, diplotype configurations and diplotype-based penetrances », *Genetics*, vol. 168, p. 2339–2348.

Larribe, F. et S. Lessard. 2008. « A composite-conditional-likelihood approach for gene mapping based on linkage disequilibrium in windows of marker loci », *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, article 27.

Larribe, F., S. Lessard, et N. J. Schork. 2002. « Gene mapping via the ancestral recombination graph », *Theoretical Population Biology*, vol. 62, p. 215–229.

Marchini, J., D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. S. Qin, H. M. Munro, G. R. Abecasis, et P. Donnelly. 2006. « A comparison of phasing algorithms for trios and unrelated individuals », *American Journal of Human Genetics*, vol. 78, p. 437–450.

McLachlan, G. J. et T. Krishnan. 2008. *The EM Algorithm and Extensions*. New-Jersey (EU) : John Wiley & Sons, Inc., 2e édition.

McPeek, M. S. et A. Strahs. 1999. « Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping », *American Journal of Human Genetics*, vol. 65, p. 858–875.

Mendel, G. 1866. « Versuche über plflanzenhybriden (expérimentation sur les végétaux) », *Verhandlungen des naturforschenden Vereines in Brünn*, vol. IV, p. 3–47.

Minichiello, M. J. et R. Durbin. 2006. « Mapping trait loci by use of inferred ancestral recombination graphs », *American Journal of Human Genetics*, vol. 79, p. 910–922.

Morris, A. P., J. C. Whittaker, et D. J. Balding. 2000. « Bayesian fine-scale mapping of disease loci, by hidden markov models », *American Journal of Human Genetics*, vol. 67, p. 155–169.

Niu, T., Z. S. Qin, X. Xu, et J. S. Liu. 2002. « Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms », *American Journal of Human Genetics*, vol. 70, p. 157–169.

Nordborg, M. 2001. « Coalescent theory ». In Balding, D. J., M. J. Bishop, et C. Cannings, éditeurs, *Handbook of Statistical Genetics*, p. 179–212, Chichester (UK). John Wiley & Sons, Inc.

Qin, Z. S., T. Niu, et J. S. Liu. 2002. « Partition-ligation–expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms », *American Journal of Human Genetics*, vol. 71, p. 1242–1247.

Stephens, M. et P. Donnelly. 2000. « Inference in molecular population genetics », *Journal of the Royal Statistical Society B*, vol. 62, p. 605–655.

———. 2001. « A new statistical method for haplotype reconstruction from population data », *American Journal of Human Genetics*, vol. 68, p. 978–989.

———. 2003. « A comparison of bayesian methods for haplotype reconstruction from population genotype data », *American Journal of Human Genetics*, vol. 73, p. 1162–1169.

Stephens, M. et P. Scheet. 2005. « Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation », *American Journal of Human Genetics*, vol. 76, p. 449–462.

Varin, C. 2008. « On composite marginal likelihoods », *Advances in Statistical Analysis*, no. 92, p. 1–28.

Xu, H., X. Wu, M. R. Spitz, et S. Shete. 2004. « Comparison of haplotype inference methods using genotypic data from unrelated individuals », *Human Heredity*, vol. 58, p. 63–68.

Zöllner, S. et J. K. Pritchard. 2005. « Coalescent-based association mapping and fine mapping of complex trait loci », *Genetics*, vol. 169, p. 1071–1092.