UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CARTOGRAPHIE DE CARACTÈRES POLYGÉNIQUES PAR LE PROCESSUS DE COALESCENCE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MYRIAM ZIOU

OCTOBRE 2017

REMERCIEMENTS

Je tiens tout d'abord à remercier Fabrice Larribe, mon directeur de recherche. Merci de m'avoir fait découvrir le monde fascinant de la statistique génétique, merci pour tes idées, ta patience, ton écoute. Merci pour les conversations enrichissantes dans ton bureau, que ce soit sur les statistiques ou sur quoique ce soit d'autre. J'ai passé 2 années fantastiques à travailler avec toi.

J'aimerais aussi remercier tous les professeurs que j'ai eus à l'UQAM, en particulier Sorana Froda, ma directrice informelle et honorifique par ses interruptions dans les échanges d'idées dans le bureau de Fabrice. Merci pour les discussions intéressantes sur tout et rien.

J'aimerais remercier Na, Delphine, Patrick, Renaud, Anthony et Abdelhakim, pour ces moments à l'université, dans le bureau ou à l'extérieur. Ça a été un réel plaisir de tous vous côtoyer.

Merci à tous mes amis et en particulier à Alexandra et Christian. Alex, merci d'avoir plus cru en moi que moi-même et d'avoir été là dans les bons moments comme dans les mauvais. Ton support incroyable et ton amitié inconditionnelle depuis 15 ans m'ont permis de venir à bout de ce mémoire entre autres choses et je t'en suis immensément reconnaissante et redevable. Christian, merci de m'avoir hébergé tous les jeudis soirs de ma dernière session où je voyageais entre Sherbrooke et Montréal. Je n'oublierai jamais ton accueil, ta générosité, tous les thalis et les soirées au ciné. Merci également d'avoir corrigé les quelques mots de latin de ce mémoire, je ne sais pas qui d'autre que toi aurait pu faire ça.

Finalement, un merci à mes parents et à mes deux magnifiques soeurs pour leur support et leur amour. Papa, merci de m'avoir transmis ta passion de la science, des mathématiques, ta curiosité. Maman, merci de toujours me dire que je suis la meilleure dans ce que je fais, même si tu ne sais pas exactement ce que c'est. Soraya et Ines, merci d'être mes soeurs, mes amies, mes modèles dans la vie.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Ce mémoire introduit une adaptation à une méthode de cartographie génétique existante à des traits causés par deux mutations. Nous commençons par énoncer certains concepts importants en génétique, puis nous présentons des méthodes de cartographie génétique existantes. Nous introduisons par la suite le processus de coalescence et un de ses dérivés important : le graphe de recombinaison ancestral. Cela nous permet de présenter la méthode sur laquelle on s'est basé ainsi que la nôtre, qui utilisent toutes deux ce type de graphe. Nous finissons par simuler des données génétiques afin de tester notre nouvelle méthode sur celles-ci en faisant varier plusieurs paramètres, ce qui nous permet de déterminer ses forces comme ses limitations.

MOTS-CLÉS : Cartographie génétique, caractère polygénique, graphe de recombinaison ancestral, maximum de vraisemblance, khi-carré.

INTRODUCTION

La génétique, une discipline encore bien jeune, connaît une évolution fulgurante depuis quelques années. À peine quinze ans se sont écoulés depuis le moment où le génome humain fut décortiqué en entier pour la première fois, en 2003, et pourtant, le premier bébé ayant le matériel génétique de trois parents a vu le jour cette année (Hamzelou (2016)). La disponibilité de données liées au bagage génétique de différentes espèces augmente et trouver des façons d'analyser ces données pour prévenir des problèmes de santé au lieu de les guérir est donc un sujet de pointe dans le monde de la recherche médicale.

Dans ce mémoire, nous adapterons *DMap*, une méthode statistique basée sur les généalogies développée par Descary (2012) visant à déterminer pour divers types de maladies héréditaires le gène qui est en cause, à un contexte où ces dites maladies sont influencées par deux gènes simultanément. Cette généralisation de *DMap* porte le nom de *DMapInteraction*.

Dans le premier chapitre, les éléments de génétique nécessaires à la compréhension de ce texte seront présentés. Le deuxième chapitre introduira quelques tests d'association communément utilisés à ce jour pour la cartographie génétique, qui est la localisation sur le génome de gènes influençant un trait. C'est dans le troisième chapitre que sera présenté le processus de coalescence, le modèle probabiliste d'évolution des populations sur lequel notre méthodologie est basée, et ses adaptations. Le quatrième chapitre introduira la méthode *DMap* ainsi que les ajustements qui lui ont été apportés ces deux dernières années pour permettre la cartographie de deux gènes à la fois, ce qui a donné naissance à *DMapInteraction*. Finalement, des

résultats obtenus avec cette dernière procédure seront présentés dans le cinquième chapitre.

CHAPITRE I

ÉLÉMENTS DE GÉNÉTIQUE

Ce chapitre a pour but d'introduire la nomenclature et d'énoncer les concepts de base de génétique utilisés dans ce mémoire. Un lecteur familier avec ce domaine est invité à passer directement au chapitre 2.

1.1 Le bagage génétique d'un individu

C'est le botaniste Johann Gregor Mendel au 19e siècle, de par son expérience sur le croisement de pois, qui est souvent considéré comme le pionnier de la génétique, cette science s'intéressant à la transmission de caractères de générations en générations. Or, la découverte de la structure en forme de double hélice de l'acide désoxyribonucléique, plus communément connu sous son acronyme ADN par James Watson et Francis Crick en 1953, fut également un moment phare dans l'avancement de cette science relativement récente. L'ADN est la source du bagage génétique de la majorité des êtres vivants, petits et grands. L'acide désoxyribonucléique est une chaîne d'une longueur avoisinant le mètre de paires de molécules qui se nomment nucléotides. Il se retrouve dans chacune de nos cellules.

Chaque nucléotide est composé d'un sucre qui porte le nom de 2'-désoxyribose, dont l'acide tire son appellation, d'un groupement phosphate ainsi que d'une base azotée pouvant être de quatre types distincts : l'adénine (A), la guanine (G), la

Figure 1.1: Représentation d'un segment d'ADN avec 20 paires de nucléotides observables.

cytosine (C) et la thymine (T). Il existe une complémentarité entre ces bases, ce qui signifie que le fait de connaître les nucléotides sur une des deux hélices composant l'ADN nous indique ce qui se retrouve sur la deuxième hélice. En effet, l'adénine est toujours couplée avec la thymine alors que la guanine l'est avec la cytosine. Cette complémentarité ainsi que la structure en double hélice permet à l'ADN de se répliquer facilement d'un point de vue biochimique, ce qui est un atout primordial pour lui étant donné qu'il se doit de se retrouver dans toutes les nouvelles cellules que le corps produit à chaque seconde.

Certaines sections de l'ADN, appelées gènes, sont responsables de la présence des caractères formant chaque individu. Il existe plus de 20 000 de ces gènes chez l'être humain, dont les longueurs peuvent aller de moins de 100 bp, ou paires de bases, à plus de 240 000 bp. L'unité de mesure équivalant à 1000 bp est le kilobase (kb) et celle équivalant à 1000 kb est la mégabase (Mb). En moyenne, un gène humain a une longueur variant entre 10 Mb et 15 Mb. Nous avons tous les mêmes gènes, mais ce qui fait en sorte que certains ont une caractéristiques associées à un gène et pas d'autres est la variante du gène qui est portée par ceux-ci. Les différentes variantes d'un gène se nomment allèles. Le génome humain, qui regroupe tous les gènes ainsi

que l'ADN qui les sépare (l'ADN intergénique), est divisé dans chaque cellule dite somatique en 23 paires de chromosomes homologues, un élément de la paire hérité du père de l'individu et l'autre de sa mère. Ces cellules sont dites diploïdes et constituent la majorité des cellules retrouvées dans le corps humain. Les cellules qui sont haploïdes, et qui ne contiennent ainsi que 23 des 46 chromosomes d'un être, sont les gamètes ou cellules sexuelles, soient l'ovule ou le spermatozoïde. Nous discuterons plus en profondeur de ces cellules dans la section suivante.

1.2 Le processus de reproduction

C'est le phénomène portant le nom de méiose qui est responsable de la création des cellules sexuelles d'un être humain, cellules ne contenant que la moitié du patrimoine génétique de celui-ci. En débutant avec une cellule diploïde, qui contient donc 23 paires de chromosomes homologues, chacun des 46 chromosomes est dupliqué, ce qui fait en sorte que la cellule contient à ce moment 92 chromosomes. On dit qu'elle est tétraploïde à cet instant. Suite à cette réplication, les chromosomes sont couplés deux par deux, le modèle étant rattaché avec sa copie en une section nommée le centromère.

Lors de la première division ou méiose I, les nouvelles paires sont toujours liées par le centromère, mais chaque pôle de la cellule attire aléatoirement soit la paire héritée de la mère de l'individu ou celle héritée par son père pour chacun des 23 types de chromosomes, dépendamment de la position des paires au moment de cette étape. Deux cellules sont donc issues de cette étape, chacune contenant 23 paires de chromosomes jumeaux. Lors de la deuxième division ou méiose II, chaque pôle de chacune de ces deux cellules attire cette fois-ci un des chromosomes jumeaux, brisant le lien au centromère, puis se divise en deux cellules identiques. Quatre cellules de 23 chromosomes sont donc le résultat de ces deux divisions successives. Étant donné que la répartition des chromosomes jumeaux lors de la

méiose I est aléatoire, il existe 2^{23} regroupements possibles de chromosomes dans les gamètes, car ceux provenant de la mère ou du père peuvent avoir été hérités de façon indépendante pour chaque position. Un exemple de l'ensemble du processus pour une cellule ayant 2 paires de chromosomes est illustré à la figure 1.2.

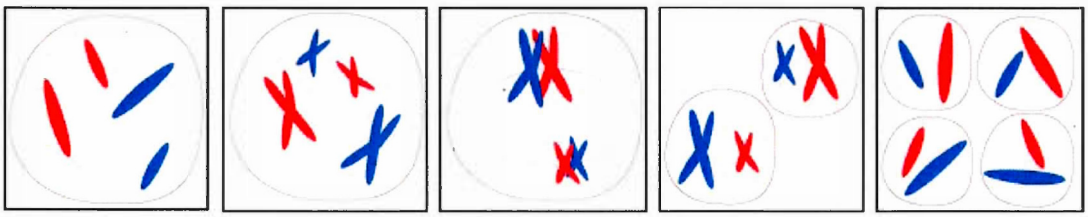

Figure 1.2: Illustration de la méiose d'une cellule diploïde contenant 2 paires de chromosomes (1^{re} paire en rouge et 2^{e} paire en bleu) donnant lieu à la création de 4 gamètes avec 2 chromosomes chacun. 1^{re} image : cellule avant la méiose. 2^{e} image : duplication des chromosomes. 3^{e} image : couplage. 4^{e} image : résultat de la méiose I. 5^{e} image : résultat de la méiose II (voir Forest (2010)).

La distribution des chromosomes lors de la méiose seule fait en sorte qu'un couple se devrait d'avoir $2^{23}+1$ enfants pour être assuré qu'il y en ait deux qui aient exactement le même bagage génétique. Par contre, dans les faits, cela en prendrait plus que ce nombre, car la création des gamètes n'est pas la seule source de diversité génétique présente chez les humains. Le phénomène de recombinaison génétique a comme conséquence qu'il y a pratiquement une infinité de cellules sexuelles qui peuvent résulter de la méiose. En effet, au début de la méiose I, lorsque les deux paires de chromosomes dupliqués sont à une très faible distance l'une de l'autre, il arrive que les branches de deux chromosomes, chacun issu d'une paire, échangent des segments d'ADN en un point se nommant chiasmata ou point de recombinaison. Cet échange a comme conséquence qu'au lieu qu'un individu transmette à

sa progéniture soit le chromosome de son père ou celui de sa mère, il transmet un hybride entre les deux. Comme les quatre chromosomes d'un type peuvent recombiner lors de la première division de la méiose, les quatre cellules haploïdes issues de la cellule diploïde peuvent contenir quatre chromosomes totalement différents pour chacun des 22 premiers types. Le phénomène de recombinaison ne se produit pas sur le 23^{e} type, la paire de chromosomes responsables de la détermination du sexe biologique de l'individu. Dans la figure 1.3, on observe deux évènements de recombinaison : un entre le premier chromosome maternel (rouge) et le premier chromosome paternel (bleu) et l'autre entre le second chromosome maternel (rouge) et le second chromosome paternel (bleu). On observe, comme résultat, quatre chromosomes distincts dont le matériel génétique est un mélange du matériel original. Le premier et le troisième chromosomes résultants sont complémentaires dans leurs matériaux rouges et bleus respectifs et il en est de même pour le deuxième et le quatrième.

Le phénomène de recombinaison s'avère bien utile pour cartographier le génome. On dit qu'il y a recombinaison entre deux points A et B sur un chromosome si le chiasmata se trouve entre eux, ce qui fait en sorte qu'ils proviennent initialement de chromosomes différents (A pourrait provenir du chromosome hérité de la mère d'un individu et B de celui hérité de son père ou vice-versa). Sous la supposition que le point où la recombinaison a lieu sur un chromosome est réparti selon une loi uniforme sur celui-ci, on peut en déduire que plus A et B sont éloignés, plus la chance qu'il y ait eu recombinaison entre eux est grande. La distance équivalente à une probabilité de 1% de recombinaison entre deux points est une unité nommée le centimorgan (cM). Déterminer les probabilités de recombinaison entre différents gènes, lorsque l'on sait qu'ils se situent sur le même chromosome, nous permet donc de schématiser l'ordre dans lequel ils se retrouvent sur ce chromosome. Par exemple, si l'on a trois positions A, B et C et que l'on sait que A a 10% de

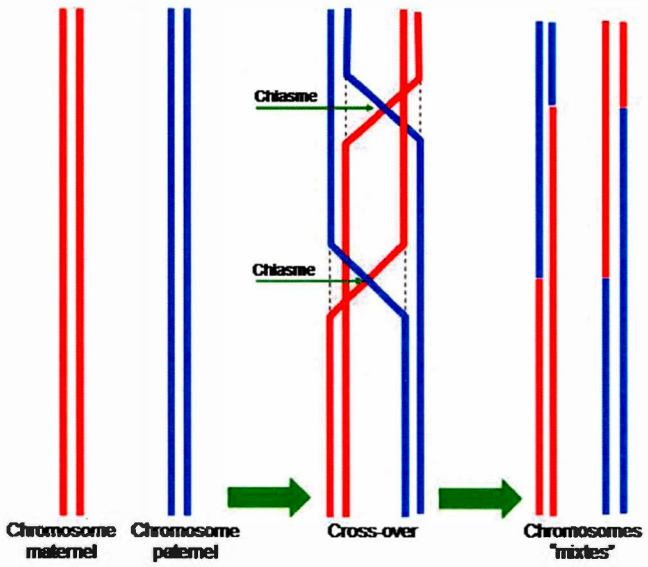

Figure 1.3: Illustration du phénomène de recombinaison entre 2 paires de chromosomes.

chance de recombiner avec B (il y a une distance de 10 cM entre A et B), 20% de recombiner avec C (il y a une distance de 20 cM entre A et C) et que B a 30% de chance de recombiner avec C (il y a une distance de 30 cM entre B et C), on peut conclure que la disposition relative qu'ont ces trois points sur le chromosome est : B, une distance de 10 cM, A, une distance de 20 cM, C.

La dernière source de diversité génétique mentionnée dans cette section est la mutation génétique. Il existe plusieurs types de transformations qui peuvent affecter notre ADN dont les causes ne sont pas encore toutes connues à ce jour. Certaines passeront inaperçues, car notre corps est en mesure de les réparer ou parce qu'elles se trouvent sur des sections de notre matériel génétique qui n'affectent pas des caractères observables, alors que d'autres auront des conséquences désastreuses. Le type de mutation qui nous intéresse dans le cadre de ce mémoire est le SNP, de

l'anglais *single nucleotid polymorphism*, le contexte dans lequel nous travaillons étant celui où une maladie d'intérêt est causée par la présence chez un individu de deux mutations distinctes de type SNP. L'appellation TIM, de l'anglais *trait influencing mutation*, est utilisée pour désigner une mutation, dans notre cas un SNP, causant un certain trait. Un SNP apparaît sur la séquence de nucléotides que forme notre génome lorsqu'une base azotée est transformée en une autre, par exemple si une base azotée de type adénine (A) est transformée en une base azotée de type cytosine (C). Un SNP peut arriver à tout moment entre la création du gamète et la fin de la vie de l'être dont ce gamète est l'origine, mais ceux qui sont une préoccupation dans les études généalogiques sont ceux qui occurent jusqu'au moment de la transmission du bagage génétique de l'être à ses descendants. Comme il est rare que, pour un nucléotide donné, il y ait plus de deux allèles possibles, on représentera dans ce mémoire une séquence génétique comme une suite de variables binaires. Un nucléotide codé 0 représentera un allèle primitif, c'est-à-dire le nucléotide qui existait originellement dans l'histoire d'une population. Pour sa part, un nucléotide codé 1 représente un allèle mutant ou SNP, donc le nucléotide en lequel le nucléotide primitif s'est transformé.

1.3 Les interactions génétiques

Dans ce mémoire, notre intérêt est dirigé envers les phénotypes, c'est-à-dire les caractères observables, qui résultent de la présence de mutations dans plus d'un gène. Certains traits sont dus à une immense quantité de gènes, ce qui les a menés à être classifiés comme traits quantitatifs, mais nous nous limiterons au cas où deux gènes sont responsables d'un trait dans ce texte.

Même à deux gènes, on peut imaginer plusieurs combinaisons produisant des effets différents. Un individu peut être porteur de 0,1 ou 2 allèles mutants pour chaque gène, donc il y a $3^2 = 9$ configurations de 2 gènes possibles. Nous discuterons plus

en profondeur de deux types de combinaisons observées dans la nature pour se donner une idée des causes et conséquences, mais il est certain qu'on aurait pu en explorer beaucoup d'autres.

La combinaison additive est un type d'interaction génétique où c'est le nombre d'allèles d'un certain type (mutant ou pas) qui détermine le caractère observé. Par exemple, il existe deux types de pigments de couleur rouge dans les fleurs d'un type de plante, *Hemerocallis*, qui est de la même famille que les lys et les jonquilles. Le premier pigment, la cyanidine, est présent si la version du gène contenue dans l'ADN est l'allèle dénoté par C et absent si, au contraire, c'est l'allèle c qui s'y trouve. Dans la même optique, le second pigment, la pélargonidine, est présent dans le cas où le gène se manifeste sous la forme P et absent lorsqu'il se retrouve sous la forme p. Comme chaque pigment est produit indépendamment, une fleur contenant 2 allèles C et 2 allèles P sera très rouge ($CCPP$) et une fleur contenant 2 allèles c et 2 allèles p sera blanche ($ccpp$). Il y a grosso-modo 3 variations intermédiaires. Les fleurs sont rouge pâle lorsqu'il y a 3 allèles C ou P et 1 allèle c ou p ($CCPp$ ou $CcPP$). Elles sont roses lorsqu'il y a 2 allèles C ou P et 2 allèle c ou p ($CCpp$, $ccPP$ ou $CcPp$). Elles sont finalement rose pâle lorsqu'il y a 1 allèle C ou P et 3 allèles c ou p ($Ccpp$ ou $ccPp$).

Le terme épistasie définit tous les types d'interaction où, contrairement à une combinaison additive, l'action d'un gène inhibe celle d'un autre. Considérons le cas de la courge dont la synthèse du pigment lui conférant sa couleur orange, la caroténoïde, nécessite une transformation par une enzyme (que nous indiquerons par A) d'un pigment vert, celui-ci ne pouvant être créé s'il y a présence d'une protéine (que nous indiquerons par B). A est générée si l'allèle G est présent au moins une fois dans le code génétique et B si l'allèle W y est présent au moins une fois. Si la courge est orange, c'est donc qu'il y avait au minimum 1 allèle G ainsi que 2 allèles w ($GGww$ ou $Ggww$). Si 2 allèles g sont présents sur le même génome

que 2 allèles w, le pigment vert est créé mais non transformé en caroténoïde, ce qui donne une couleur verte à la courge ($ggww$). Dans le cas où on avait au moins un allèle W, peu importe si la courge est porteuse d'allèles G ou g, le pigment vert qui est primordial à la fabrication du pigment orange est inexistant, la courge est ainsi blanche ($WWGg$, $WWGG$, $WWgg$, $WwGg$, $WwGG$ ou $Wwgg$). On dit que W et G sont des gènes épistatiques, car W masque G dans notre cas.

Il est mentionné par Phillips (2008) que des maladies graves et bien présentes dans la société d'aujourd'hui comme le diabète, l'insuffisance coronarienne et l'autisme ont déjà été associées à des gènes épistatiques à travers des études d'association. Développer des méthodes afin de faire de la cartographie génétique à deux gènes ou plus est donc primordial afin de permettre de mieux comprendre et étudier les causes de maladies de ce type.

CHAPITRE II

CARTOGRAPHIE GÉNÉTIQUE

Lorsque l'on cherche à déterminer la position d'un gène qui est responsable d'une maladie ou d'un autre caractère que l'on désire étudier, les méthodes qui s'offrent à nous sont très nombreuses et le choix de celles à privilégier dépend de plusieurs facteurs. Le type d'étude mené et le type de phénotype recherché n'en sont que deux. Nous ne présenterons dans ce chapitre que deux des méthodes les plus classiques dans les études cas-témoins portant sur les phénotypes binaires, qui sont le type d'études traitées dans le cadre de ce mémoire. Elles seront développées toutes les deux pour le cadre où l'on cherche à mesurer l'association entre le phénotype et un marqueur sur le génome, dans un premier temps, puis, pour celui où nous pensons que deux marqueurs sont causaux au phénotype étudié, dans un deuxième temps. La dernière section de ce chapitre correspondra à des applications de ces méthodes à des données simulées. Préalablement à la présentation des méthodes, nous introduirons le déséquilibre de liaison, ce phénomène génétique qui permet de ne pas avoir à séquencer le génome en entier pour localiser de façon efficace les gènes recherchés.

2.1 Le déséquilibre de liaison

Lorsque nous considérons deux loci sur le génome, les allèles présents à ces loci peuvent s'être retrouvés transmis ensemble au cours de l'histoire plus souvent que

ce ne serait le cas par le simple fruit du hasard. La mesure de dépendance entre des allèles à des positions distinctes sur un même haplotype se nomme le déséquilibre de liaison, souvent désigné dans la littérature par le diminutif LD, de l'anglais *linkage desiquilibrium.* Considérons deux marqueurs où seulement deux allèles peuvent exister pour chacun, nommons ces allèles A et a au premier marqueur et B et b au second. Désignons par D_{AB} la différence entre la fréquence des haplotypes portant à la fois les versions A et B des allèles et la fréquence s'il y avait indépendance totale entre le fait de porter A et le fait de porter B, qui correspond au produit des fréquences de ces deux allèles, tel que formulé dans l'équation (2.1) :

$$\begin{aligned} D_{AB} &= p_{AB} - p_A p_B \qquad (2.1) \\ &= p_{AB} - (p_{AB} + p_{Ab})(p_{AB} + p_{aB}) \\ &= p_{AB} - p_{AB}^2 - p_{AB}p_{aB} - p_{Ab}p_{AB} - p_{Ab}p_{aB} \\ &= p_{AB}(1 - p_{AB} - p_{aB} - p_{Ab}) - p_{Ab}p_{aB} \\ &= p_{AB}p_{ab} - p_{Ab}p_{aB}. \qquad (2.2) \end{aligned}$$

L'équation (2.2) correspond à cette mesure réécrite en terme des 4 fréquences conjointes d'allèles possibles. En utilisant le même type de démarche, il est possible de montrer que $D_{ab} = D_{AB}$ et que $D_{Ab} = D_{aB} = -D_{AB}$. L'utilisation d'une seule variable, D, dans un contexte où chacun des deux marqueurs n'a que deux allèles possibles est donc la norme.

Étant donné que le phénomène de recombinaison a un effet majeur sur la présence d'équilibre ou de déséquilibre de liaison, car plus il y a de chance d'avoir de recombinaisons entre deux allèles, moins les chances qu'ils soient transmis conjointement lors de la reproduction sont grandes, la connaissance de la fraction de recombinaison, θ, contribue à déterminer la variation de D dans le temps. Le paramètre θ exprime la probabilité qu'il y ait recombinaison entre deux allèles lors d'une

méiose. Elle peut donc prendre des valeurs variant entre 0 et 0.5, 0 correspondant au cas où des allèles sont tellement rapprochés qu'il est impossible qu'il y ait recombinaison entre eux et 0.5 à des allèles indépendants. θ est généralement facile à estimer lorsqu'on connaît la distance entre les deux allèles ainsi que la région du chromosome où on se trouve. Donc, lorsqu'on connaît la valeur de la mesure de LD à un moment dans le temps, D_t, il est possible de déterminer celle-ci une génération plus tard comme suit :

$$D_{t+1} = (1 - \theta)D_t. \tag{2.3}$$

L'équation (2.3) correspond à l'espérance de la mesure de LD. D'une part, cette dernière peut prendre la valeur 0 s'il y a eu recombinaison entre les deux marqueurs, car ils sont transmis de façon indépendante au descendant dans un tel cas. D'autre part, si un tel évènement n'a pas lieu, la mesure de dépendance entre les deux marqueurs au temps t, D_t, est inchangée. La figure 2.1 exprime la diminution d'une mesure de LD initiale $D = 0.5$ en fonction du temps pour trois valeurs de fractions de recombinaison différentes. On remarque que plus il y a de chances de recombiner entre deux allèles, plus on va se rapprocher de l'équilibre de liaison rapidement.

Étant donné que l'interprétation et la comparaison des mesures de différence D n'est pas évidente, car cette métrique est très influencée par les fréquences alléliques des marqueurs concernés, plusieurs statistiques basées sur celle-ci ont été développées.

En déterminant l'étendue de D comme ceci

$$D_{\min} = \max\{-p_A p_B, -p_a p_b\}, \tag{2.4}$$

$$D_{\max} = \min\{p_A p_b, p_B p_a\}, \tag{2.5}$$

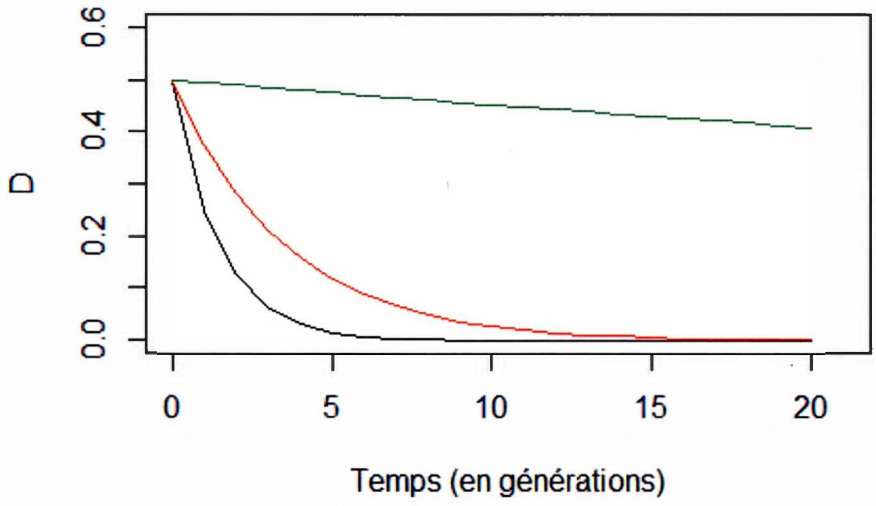

Figure 2.1: Variation d'une mesure de LD de 0.5 en fonction du temps. Pour θ=0.01 (vert), θ=0.25 (rouge) et θ=0.5 (noir).

le D', introduit par Lewontin (1964), se calcule ainsi :

$$D' = \frac{D}{\max\{|D_{\min}|, |D_{\max}|\}}. \tag{2.6}$$

La valeur absolue de cette statistique peut varier entre 0 et 1, où 0 correspond à un équilibre ou indépendance parfaite entre les allèles et 1 à un déséquilibre complet, caractérisé par la disparition d'au moins une des quatre combinaisons possibles (AB, Ab, aB ou ab), alors que p_A, p_a, p_B et p_b sont tous non-nuls.

Une deuxième statistique importante basée sur D est

$$\delta_A = p_A + \frac{D}{p_B}. \tag{2.7}$$

En substituant dans (2.7) D tel que défini dans (2.1), on montre qu'elle définit la probabilité d'un haplotype d'être porteur de l'allèle A sachant qu'il est porteur de

l'allèle B :

$$\begin{aligned} \delta_A &= p_A + \frac{p_{AB} - p_A p_B}{p_B} \\ &= p_A + \frac{p_{AB}}{p_B} - p_A. \\ &= \frac{p_{AB}}{p_B} \end{aligned} \tag{2.8}$$

Une représentation graphique de la statistique $|D'|$ quantifiant le déséquilibre de liaison est présentée à la figure 2.3. Au lieu d'avoir deux axes perpendiculaires identiques où le croisement d'une abcisse x et une ordonnée y nous donnerait la valeur de $|D'|$ pour les marqueurs représentés par (x, y), nous utilisons un type de graphique ne comportant qu'un axe contenant tous les marqueurs d'une séquence donnée. Pour comprendre comment le lire, observons la figure 2.2. On a une séquence où il y a 4 marqueurs, $M1$, $M2$, $M3$ et $M4$. Pour déterminer la valeur de $|D'|$ pour $M1$ et $M4$, il faut regarder la couleur du point à la jonction de la diagonale issue de $M1$ et celle issue de $M4$. Plus on se rapproche du noir, plus les marqueurs sont liés alors que si on tend vers le blanc, les marqueurs se rapprochent de l'équilibre de liaison. En d'autres mots, un point noir définit $|D'| = 1$, un point blanc définit $|D'| = 0$ et un spectre de gris très foncé à très clair définit les valeurs intermédiaires.

Dans la figure 2.3, on sépare un échantillon de 1000 individus simulés en deux sous-échantillons : celui des malades (500 individus) et celui des non-malades (500 individus). On calcule ensuite $|D'|$ pour tous les marqueurs de notre séquence dans les deux sous-échantillons indépendamment. La séquence est d'une longueur de 999.5 kb (999 500 paires de bases) et c'est la ligne blanche pointillée qui fait figure de l'unique axe. Celui-ci n'est pas gradué pour des raisons esthétiques, mais le haut représente le début de la séquence et le bas, la fin. Les témoins sont illustrés à gauche de l'axe et les cas à droite. Les lignes rouges sont superposées sur les diagonales provenant des deux marqueurs causant la maladie. La raison

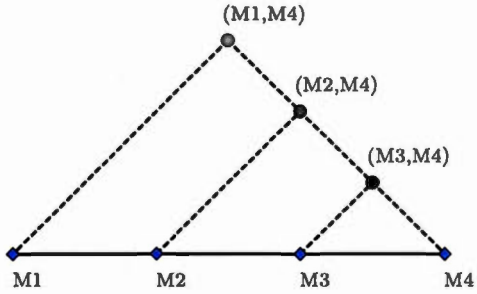

Figure 2.2: Visualisation de la technique pour lire un graphique de LD comme celui de la figure 2.3.

pour laquelle la figure n'est pas un losange est que nous n'avons conservé que les paires de marqueurs pour lesquels la distance était en deçà de 250 kb et ceux dont les allèles de fréquence mineure (ou MAF, la fréquence de l'allèle le moins présent des deux à un marqueur, on se souvient que nos marqueurs sont tous biallélíques) pour les deux marqueurs étaient supérieurs à 0.01.

On remarque que les couleurs sont généralement plus foncées à droite, chez les cas, qu'à gauche, chez les témoins. En effet, étant donné que la maladie est génétique et que les malades partagent donc des mutations causales qui sont rares, ils sont probablement plus liés généalogiquement que les non-malades qui proviennent d'un bassin de population plus vaste et hétérogène. On remarque également une multitude de petits triangles très foncés des deux côtés de l'axe. Chez les cas comme chez les témoins, on peut être porté à croire que c'est entre ces triangles qu'il y a eu des évènements de recombinaison dans l'histoire chez certains des individus. En effet, on a mentionné plus haut que c'était les recombinaisons qui faisaient en sorte que les marqueurs étaient transmis de façon indépendante et que plus deux marqueurs étaient éloignés, plus les chances qu'ils recombinent étaient grandes. Donc des marqueurs très près entre lesquels il n'y a eu recombinaison chez presque

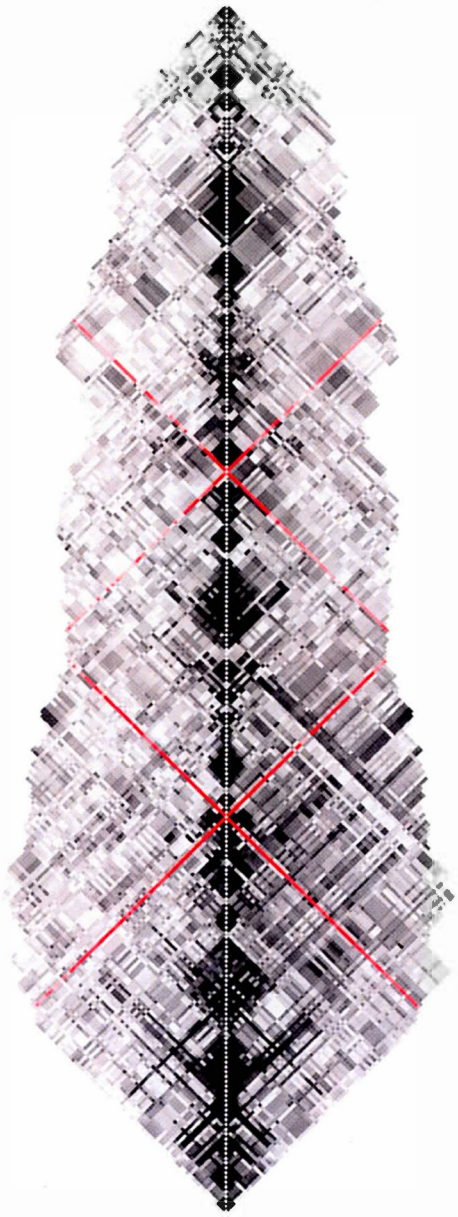

Figure 2.3: Mesure de $|D'|$ le long d'une séquence. Les témoins sont à gauche de l'axe en pointillés et les cas à droite.

personne ont une dépendance quasi parfaite. À l'inverse, plus deux marqueurs sont loins, moins en moyenne, la dépendance est grande d'où la gradation habituelle du foncé vers le pâle lorsqu'on s'éloigne de l'axe. Par contre, si on porte bien attention, on voit que la distance n'est pas tout, car on a certains marqueurs très éloignés qui ont un $|D'|$ élevé et d'autres qui sont à proximité mais dont le $|D'|$ est faible. Ces variations sont entre autres influencées par les MAF, la dépendance entre les individus dans l'échantillon sélectionné ainsi que sa taille.

Le déséquilibre de liaison est un phénomène primordial en cartographie génétique, car c'est celui qui permet de déterminer des blocs d'haplotypes (une traduction libre de *haplotype blocks*), c'est-à-dire des régions où les marqueurs sont très fortement associés les uns aux autres (comme les triangles noirs qu'on voyait dans la figure 2.3). Ceci permet de pouvoir détecter un ou des marqueurs causaux non pas directement, mais par leur lien avec des marqueurs pour lesquels on trouve une association avec un phénotype (des méthodes afin d'y arriver sont amenées dans les sections suivantes), ce qui est indispensable quand le marqueur recherché n'est pas séquencé.

2.2 L'association par le χ^2

Il existe aujourd'hui une multitude de façons de déterminer si un marqueur génétique et un phénotype binaire comme une maladie d'intérêt sont corrélés, que le marqueur cause directement le phénotype ou non. Une de celles-ci est l'utilisation d'un test simple ancien mais encore très courant considérant son efficacité : le test d'indépendance de deux variables qualitatives. Les variations qui peuvent être apportées à ce test sont nombreuses.

2.2.1 L'association avec un unique marqueur

Une façon intuitive de mesurer l'association entre un marqueur M diallélique pouvant prendre les formes A ou a et le fait d'être un cas (malade) ou un témoin (non-malade) est de se servir d'un tableau de contingence comme le tableau 2.1.

	Cas	Témoins	Total
AA	$n_{cas,AA}$	$n_{tem,AA}$	n_{AA}
Aa	$n_{cas,Aa}$	$n_{tem,Aa}$	n_{Aa}
aa	$n_{cas,aa}$	$n_{tem,aa}$	n_{aa}
Total	n_{cas}	n_{tem}	n

Tableau 2.1: Tableau de contingence de la présence d'une maladie et du génotype au marqueur M pour n individus.

On rappelle que la statistique du khi-deux se calcule en déterminant, dans un premier temps, chacune des fréquences espérées E_{ij} ($i \in \{1,2,3\}$, $j \in \{1,2\}$) par le produit de le i^e ligne et j^e colonne divisée par n, puis en faisant le calcul :

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{2.9}$$

Sous l'hypothèse nulle d'indépendance, elle est distribuée selon une loi du χ^2 avec $(3-1)(2-1) = 2$ degrés de liberté.

Outre ce tableau de contingence général, il serait possible d'en construire d'autres si on connaît a priori des informations sur notre phénotype. Par exemple, dans le cas d'une maladie récessive pour l'allèle A (que l'on considère comme la version mutée du gène), où il est donc nécéssaire d'avoir le génotype AA pour être atteint, on peut regrouper les génotypes aa et Aa comme une seule sous-population. Au

contraire, si la maladie est dominante pour ce même allèle A, et donc qu'un seul allèle A suffit pour être affecté, c'est les individus porteurs de Aa et AA qui peuvent être réunis dans une même catégorie. Comme pour ces deux tableaux, une ligne est retirée, les degrés de liberté des statistiques du χ^2 sont maintenant $(2-1)(2-1) = 1$ dans les deux cas.

2.2.2 L'association avec deux marqueurs

Parfois, on peut avoir comme hypothèse que l'existence d'un phénotype d'intérêt ne peut être expliquée que par la connaissance d'un seul marqueur causal. Cela peut être occasionné par de multiples causes, l'une d'entre elles étant celle à laquelle on s'intéresse dans ce travail : l'existence d'un second gène nécessaire au développement du caractère d'importance.

Le tableau de contingence du test d'indépendance du khi-deux pour ce type de maladies en est un de dimension 9x2, étant donné que bien qu'il y ait toujours deux statuts, cas et témoins, les possibilités de couples de génotypes sont maintenant $3^2 = 9$, comme illustré dans le tableau 2.2.

La statistique à calculer est donc maintenant :

$$\chi^2 = \sum_{i=1}^{9}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \tag{2.10}$$

où les E_{ij} correspondent encore au total de la ligne fois le total de la colonne divisé par la taille d'échantillon n. Sous l'hypothèse nulle d'indépendance, elle est distribuée selon une loi du χ^2 avec $(9-1)(2-1) = 8$ degrés de liberté. Le problème majeur avec un tableau de cette ampleur est qu'il peut être ardu et coûteux d'obtenir la taille d'échantillon nécessaire pour que l'approximation soit acceptable, considérant qu'il est relativement fréquent d'obtenir des cases vides et même des totaux nuls lorsque certains allèles ou le phénotype sont rares. Nous verrons dans le chapitre 4 de ce mémoire, comment nous procédons pour tenter

	Cas	Témoins	Total
$AABB$	$n_{cas,AABB}$	$n_{tem,AABB}$	n_{AABB}
$AABb$	$n_{cas,AABb}$	$n_{tem,AABb}$	n_{AABb}
$AAbb$	$n_{cas,AAbb}$	$n_{tem,AAbb}$	n_{AAbb}
$AaBB$	$n_{cas,AaBB}$	$n_{tem,AaBB}$	n_{AaBB}
$AaBb$	$n_{cas,AaBb}$	$n_{tem,AaBb}$	n_{AaBb}
$Aabb$	$n_{cas,Aabb}$	$n_{tem,Aabb}$	n_{Aabb}
$aaBB$	$n_{cas,aaBB}$	$n_{tem,aaBB}$	n_{aaBB}
$aaBb$	$n_{cas,aaBb}$	$n_{tem,aaBb}$	n_{aaBb}
$aabb$	$n_{cas,aabb}$	$n_{tem,aabb}$	n_{aabb}
Total	n_{cas}	n_{tem}	n

Tableau 2.2: Tableau de contingence de la présence d'une maladie et des génotypes aux marqueurs M_1 et M_2 pour n individus.

de contourner ces difficultés dans le cadre des tests d'indépendance effectués par notre méthode.

Les forces principales des tests d'association par le khi-deux sont qu'ils sont simples et très rapides à réaliser informatiquement. Ce dernier est un avantage à ne pas négliger quand on a beaucoup de marqueurs à tester comme dans les études de types GWAS, ou *Genome Wide Association Studies*, qui cartographient le génome complet. Ceci est d'autant plus important quand on doit cartographier une maladie causée par plus d'une mutation. Un autre avantage intéressant de ces tests d'association est qu'ils ne requièrent aucune connaissance préalable du modèle génétique de la maladie. Par contre, tout n'étant pas blanc ou noir, ces méthodes ont aussi des inconvénients. Le plus notable est que les conditions des tests du chi-deux exigent un échantillon indépendant, ce qui est parfois difficile à obtenir pour

des maladies rares à fort caractère génétique, car les individus affectés peuvent être reliés. Aussi, ils nécessitent dans le cas du tableau 9x2 des tailles d'échantillon très grandes lorsque la maladie et les mutations sont trop rares pour avoir assez d'individus dans chaque case, ce qui n'est pas toujours une condition facile d'application. De plus, étant donné le très grand nombre de tests statistiques à effectuer, trouver une valeur critique adéquate pour s'assurer une bonne capacité de détection tout en évitant les faux positifs, c'est-à-dire de détecter une association à la maladie pour des marqueurs pour lesquels on ne le devrait pas, n'est pas une mince tâche. Enfin, les marqueurs sont liés spatialement entre eux, comme on l'a vu dans la section 2.1 sur le déséquilibre de liaison. Donc, un nombre important de tests qui ne sont pas indépendants est effectué.

2.3 La régression logistique

Étant donné que la présence de maladie ou non chez un individu est un trait dichotomique, la régression logistique s'avère être un outil précieux dans un contexte de cartographie génétique.

2.3.1 L'association avec un seul marqueur

Une façon d'utiliser cet outil est par la fonction de lien logit entre un prédicteur linéaire η_i et la pénétrance f_i, qui correspond à la probabilité d'être un cas connaissant le nombre d'allèles mutants i portés à un certain marqueur M (considérant que l'allèle muté est A, $i = 0$ si le génotype d'un individu est aa, $i = 1$ s'il est Aa et $i = 2$ s'il est AA) :

$$\eta_i = ln\left(\frac{f_i}{1 - f_i}\right). \tag{2.11}$$

η_i peut être interprété comme le logarithme du rapport entre la chance d'être malade et celle d'entre en santé si on a i allèles mutants au marqueur M. Ainsi, si les chances d'être malade sont plus grandes que celles de ne pas l'être sachant i, η_i

sera positif et il sera négatif dans le cas contraire. Une façon classique de modéliser ce prédicteur linéaire pour un individu j est d'utiliser un modèle logistique de la forme :

$$\eta_j = \beta_0 + \beta_1 z_j, \tag{2.12}$$

où la variable z_j en est une catégorielle qui peut, par exemple, être codée par le nombre d'allèles mutants au marqueur M (z_j est donc égal au $i \in \{0,1,2\}$ mentionné plus haut). Une autre façon rencontrée fréquemment dans la littérature, dont dans Morris et Cardon (2007), de modéliser η_j est par la relation :

$$\eta_j = \beta_0 + \beta_A z_{(A)j} + \beta_D z_{(D)j}. \tag{2.13}$$

Les variables $z_{(A)j}$ et $z_{(D)j}$ sont des variables catégorielles prenant respectivement les valeurs -1,0,1 et 0,1,0 pour les valeurs de i correspondant à 0, 1 et 2. β_A mesure donc l'effet additif, positif ou négatif, du nombre d'allèles mutants sur la maladie, alors que β_D quantifie un effet de dominance de l'allèle a sur l'allèle A. En effet, si β_D est non-nul, cela indique que le fait d'être homozygote a un effet positif ou négatif sur la pénétrance si on se compare au fait d'être hétérozygote.

On rappelle qu'une statistique de rapport de log-vraisemblance Λ correspond au double de la différence entre les logarithmes des vraisemblances de deux modèles estimés emboîtés, le premier étant celui où le plus de paramètres sont estimés. Λ suit approximativement une loi du χ^2 où le nombre de degrés de liberté correspond à la différence du nombre de paramètres estimés entre les deux. Pour ce qui est du modèle (2.12), on peut calculer :

$$\Lambda = 2[ln(L(\hat{\beta}_0, \hat{\beta}_1)) - ln(L(\hat{\beta}_0, \beta_1 = 0))], \tag{2.14}$$

qui suit approximativement une loi du χ^2 à 2 degrés de liberté, car z_j pouvant prendre 3 modalités, il faut estimer 2 paramètres pour la représenter et un $\beta_1 = 0$,

qui signifie que le marqueur M n'a pas d'effet sur le phénotype nécessite que les deux soient nuls.

Pour ce qui est du modèle (2.13), on peut également tester l'hypothèse de nullité des deux paramètres estimés :

$$\Lambda = 2[ln(L(\hat{\beta}_0, \hat{\beta_A}, \hat{\beta_D})) - ln(L(\hat{\beta}_0, \beta_A = \beta_D = 0))] \tag{2.15}$$

qui nous permet de déterminer à l'aide d'une règle de décision basée sur le khi-carré si le marqueur M a un effet quelconque sur la probabilité d'être malade ou si tous les $\eta_i = \beta_0$, ce qui fait que $f_0 = f_1 = f_2$. Il ferait plus de sens dans cette modélisation de tester également la nullité d'un seul des paramètres, pour ne déterminer que l'existence d'un effet additif ou d'un effet de dominance.

2.3.2 L'association avec deux marqueurs

Quand on tente de déterminer si deux marqueurs dialléliques M_1 et M_2, définis par les indices i et k peuvent être conjointement mis en cause pour la présence d'un phénotype chez un individu j, on définit le prédicteur linéaire η_{ik} comme :

$$\eta_{ik} = ln\left(\frac{f_{ik}}{1 - f_{ik}}\right), \tag{2.16}$$

où la pénétrance f_{ik} correspond à la probabilité d'être malade sachant le nombre d'allèles mutants i et k portés aux marqueurs M_1 et M_2 (i et k peuvent donc chacun valoir 0,1 ou 2).

La généralisation du modèle (2.12) à un contexte où on tente de déterminer si deux gènes aux marqueurs dialléliques M_1 et M_2, définis par les indices i et k, peuvent être conjointement mis en cause pour la présence d'un phénotype chez un individu j est

$$\eta_j = \beta_0 + \beta_i z_{ij} + \beta_k z_{kj} + \beta_{ik} z_{ij} z_{kj}. \tag{2.17}$$

La variable z_{ij} peut être codifiée de diverses façons. Une d'entre elles, fréquemment utilisée, est d'assigner $z_{ij} = 1$ si l'individu j a 2 allèles mutants au premier marqueur M_1, $z_{ij} = 2$ s'il en a 1 et $z_{ij} = 3$ s'il n'en a pas. La même modélisation est utilisée pour z_{kj} en fonction du nombre d'allèles mutants au deuxième marqueur M_2.

Wan et al. (2013) introduisent deux tests d'hypothèses majeurs pouvant être appliqués à cette régression. Le test d'interaction consiste à calculer

$$\Lambda = 2[ln(L(\hat{\beta}_0, \hat{\beta}_i, \hat{\beta}_k, \hat{\beta}_{ik})) - ln(L(\hat{\beta}_0, \hat{\beta}_i, \hat{\beta}_k, \beta_{ik} = 0))] \qquad (2.18)$$

afin de déterminer s'il y a une interaction entre les deux marqueurs ou s'ils contribuent peut-être tous les deux à la maladie, mais de façon indépendante et additive. Dans ce test, Λ suit approximativement une loi du χ^2 à 4 degrés de liberté.

Le second test mentionné dans l'article est le test d'association complète qui compare le modèle où aucun des deux marqueurs n'est impliqué dans la présence de la maladie avec le modèle complet de l'équation (2.17) en calculant

$$\Lambda = 2[ln(L(\hat{\beta}_0, \hat{\beta}_i, \hat{\beta}_k, \hat{\beta}_{ik})) - ln(L(\hat{\beta}_0, \beta_i = \beta_k = \beta_{ik} = 0))], \qquad (2.19)$$

où Λ suit approximativement une loi du χ^2 à 8 degrés de liberté. Une sélection de marqueurs qui ont un effet significatif est faite préalablement pour ne pas détecter des faux positifs lorsqu'on teste des couples où un marqueur est très impliqué dans la maladie et un qui ne l'est pas du tout. Par contre, le risque de premier espèce déterminé pour ce test préliminaire doit être assez faible pour ne pas rejetter des marqueurs qui pourraient être impliqués.

En ce qui a trait au modèle (2.13), son adaptation à la nouvelle problématique, que l'on peut également retrouver dans Morris et Cardon (2007), est pour un individu j :

$$\eta_j = \beta_0 + \beta_{Ak} z_{(Ak)j} + \beta_{Dk} z_{(Dk)j} + \beta_{Al} z_{(Al)j} + \beta_{Dl} z_{(Dl)j} + \beta_{AAkl} z_{(Ak)j} z_{(Al)j}$$

$$+\beta_{ADkl}z_{(Ak)j}z_{(Dl)j} + \beta_{DAkl}z_{(Dk)j}z_{(Al)j} + \beta_{DDkl}z_{(Dk)j}z_{(Dl)j}. \qquad (2.20)$$

On peut ainsi déterminer les effets individuels de chacun des deux gènes en étudiant β_{Ak}, β_{Dk}, β_{Al} et β_{Dl}, mais aussi mesurer l'interaction entre les effets additifs et dominants de chacun à l'aide de β_{AAkl}, β_{ADkl}, β_{DAkl} et β_{DDkl}. Le tableau 2.3 résume le signe de chacun des coefficients estimé conditionnellement aux deux génotypes présents chez un individu, lorsque l'on adopte le même choix de variables que celui défini pour le cas monogénique.

	$z_{(Ak)}$	$z_{(Dk)}$	$z_{(Al)}$	$z_{(Dl)}$	$z_{(Ak)}z_{(Al)}$	$z_{(Ak)}z_{(Dl)}$	$z_{(Dk)}z_{(Al)}$	$z_{(Dk)}z_{(Dl)}$
$AABB$	-1	0	-1	0	1	0	0	0
$AABb$	-1	0	0	1	0	-1	0	0
$AAbb$	-1	0	1	0	-1	0	0	0
$AaBB$	0	1	-1	0	0	0	-1	0
$AaBb$	0	1	0	1	0	0	0	1
$Aabb$	0	1	1	0	0	0	1	0
$aaBB$	1	0	-1	0	-1	0	0	0
$aaBb$	1	0	0	1	0	1	0	0
$aabb$	1	0	1	0	1	0	0	0

Tableau 2.3: Valeurs des covariables associées aux coefficients β du modèle de l'équation (2.20) pour tous les couples de génotypes possibles aux marqueurs M_1 et M_2.

Pour tester simultanément si A ou B ont un effet quelconque sur la maladie, on calcule une statistique de rapport de log-vraisemblance similaire à celle en (2.19), qui suit approximativement une loi du χ^2 où le nombre de degrés de liberté est 8 :

$$\Lambda = 2[ln(L(\hat{\beta}_0, \hat{\beta}_{Ak}, \hat{\beta}_{Dk}, \hat{\beta}_{Al}, \hat{\beta}_{Dl}, \hat{\beta}_{AAkl}, \hat{\beta}_{ADkl}, \hat{\beta}_{DAkl}, \hat{\beta}_{DDkl}))$$

$$-ln(L(\hat{\beta}_0, \beta_{Ak} = \beta_{Dk} = \beta_{Al} = \beta_{Dl} = \beta_{AAkl} = \beta_{ADkl} = \beta_{DAkl} = \beta_{DDkl} = 0))]. \quad (2.21)$$

La prochaine statistique, analogue à celle en (2.18), a l'utilité de déterminer s'il existe un effet d'épistasie entre les deux gènes, et donc que leur effet commun sur le phénotype ne se résume pas à la simple addition de leurs effets individuels :

$$\begin{aligned} \Lambda = 2[&ln(L(\hat{\beta}_0, \hat{\beta}_{Ak}, \hat{\beta}_{Dk}, \hat{\beta}_{Al}, \hat{\beta}_{Dl}, \hat{\beta}_{AAkl}, \hat{\beta}_{ADkl}, \hat{\beta}_{DAkl}, \hat{\beta}_{DDkl})) \\ &-ln(L(\hat{\beta}_0, \hat{\beta}_{Ak}, \hat{\beta}_{Dk}, \hat{\beta}_{Al}, \hat{\beta}_{Dl}, \beta_{AAkl} = \beta_{ADkl} = \beta_{DAkl} \\ &= \beta_{DDkl} = 0))]. \end{aligned} \quad (2.22)$$

Ce dernier Λ suit approximativement une loi du χ^2 de 4 degrés de liberté. Si on cherche à mieux comprendre la maladie à laquelle on fait face et son caractère dominant ou non, il serait également possible de tester chacun des coefficients individuellement. Pour plus d'information sur cette seconde modélisation par la régression logistique, on peut se référer à Morris et Cardon (2007).

2.4 Comparaison de méthodes d'association

Afin de mettre en pratique certaines des méthodologies introduites dans ce chapitre, nous avons décidé de simuler une population de 10 000 individus, d'en extraire trois échantillons selon des modèles génétiques différents et d'illustrer les conclusions auxquelles on arrive à l'aide des méthodes d'association simple par le khi-carré, d'association avec deux marqueurs par le khi-carré et d'une des régressions logistiques multiples. Les tests comme les graphiques dans cette section ont été faits avec le logiciel R.

Pour la méthode d'association simple par le khi-carré, on utilise le test associé au tableau de contingence 2.1, où les lignes correspondent au génotype pour chacun des marqueurs de nos ensembles de données et les colonnes au statut malade

ou non-malade des individus. Les premiers graphiques pour chacune des figures correspondent donc à l'inverse du logarithme de la p-valeur pour la statistique du χ^2 calculée pour chacun des marqueurs. Les droites verticales rouges représentent les positions des deux marqueurs causaux.

Pour la méthode d'association avec 2 marqueurs par le khi-carré, on utilise le test le plus général, associé au tableau de contingence 2.2. Les lignes correspondent aux 9 possibilités de combinaison de génotypes pour chacun des couples de marqueurs de nos ensembles de données et les colonnes au statut malade ou non-malade des individus. Les deuxièmes graphiques pour chacune des figures correspondent donc à l'inverse du logarithme de la p-valeur pour la statistique du χ^2 calculée pour chacun des couples de marqueurs. Les droites noires pleines définissent les deux marqueurs causaux et les droites noires pointillées l'ensemble de 2 marqueurs pour lequel nous avons trouvé la plus forte association avec le phénotype.

Finalement, pour la régression logistique, nous avons choisi le test d'association complète introduit par le modèle (2.17). Par contre, nous n'avons pas fait l'étape initiale de présélection de marqueurs significatifs. Dans cette section, nous ne cherchons pas à faire un test d'hypothèses dans les règles de l'art avec un seuil fixé, la détermination d'un seuil lors de tests multiples comme ceux-ci étant une problématique pas toujours évidente en soi. Notre but est plutôt de faire une détection des marqueurs avec la plus forte association. Les troisièmes graphiques pour chacune des figures correspondent ainsi à l'inverse du logarithme de la p-valeur pour la statistique Λ définie dans l'équation (2.19) et déterminée pour chacun des couples de marqueurs. Les droites pleines et pointillées ont les mêmes significations que pour les deuxièmes graphiques.

Le premier échantillon en est un de 500 cas et 500 témoins. La maladie créée dans nos données est telle qu'on a une chance de 99% d'être affecté si on a au moins un

allèle causal à chacun des deux marqueurs. A et B étant les allèles causaux, on est donc presque certain d'être un cas si on est $AaBb$, $AABb$, $AaBB$ ou $AABB$.

Le deuxième échantillon comporte 400 cas et 400 témoins. Dans cet échantillon, c'est le nombre d'allèles causaux portés qui détermine notre chance d'être malade et non pas leur provenance. On a 60% de chance d'être malade si on en a 2 (donc si on est $AaBb$, $AAbb$ ou $aaBB$), 90% de chance d'être malade si on en a 3 (donc si on est $AABb$ ou $AaBB$) et 99% de chance d'être malade si on en a 4 (si on est $AABB$).

Le troisième échantillon compte également 400 cas et 400 témoins. Le modèle utilisé pour déterminer les cas et les témoins en est un où M_1 a un impact beaucoup plus grand sur le développement de la maladie que M_2. On a respectivement 60%, 70%, 80% et 90% de chance d'être malade si on est $Aabb$, $AaBB$, $AAbb$ et $AABb$ alors qu'on n'a que 2% si on est $aaBb$ et 5% si on est $aaBB$.

Les figures 2.4 et 2.5 des pages suivantes illustrent toutes deux les résultats obtenus à partir du premier échantillon. La différence entre les deux est que les vrais TIMs qui, rappelons-le, sont les mutations partiellement ou totalement responsables de la présence d'un trait observé, font partie des données utilisées dans la figure 2.5, alors qu'ils sont exclus de celles utilisées dans la figure 2.4. On repose donc sur l'espoir que des marqueurs en fort déséquilibre de liaison avec nos marqueurs causaux soient aussi associés fortement avec la maladie, ce qui aiderait à la détection des marqueurs recherchés. Les figures 2.6 et 2.7 sont dans la même situation pour le deuxième échantillon et les figures 2.8 et 2.9 pour le troisième.

Si on observe la figure 2.4, on remarque que déjà de façon unidimensionnelle, on observe des tendances d'association plus fortes autour de nos deux marqueurs. La p-valeur minimale de toutes nos données se trouve sur le marqueur le plus près du deuxième TIM, mais la détection semble un peu plus difficile pour le

premier où des marqueurs qui ne sont pas les plus près de celui-ci semblent être en plus forte association avec la maladie que celui le plus près de la première ligne rouge. Ceci se répercute dans les deux graphiques plus bas, où les couples avec l'association la plus forte sont dans les deux cas un où le deuxième marqueur causal est contenu, mais pas le premier. Le fait d'avoir accès à l'information sur nos marqueurs causaux dans cette situation change réellement la donne, comme on peut le voir la détection exacte dans la figure 2.5. On remarque également que pour le premier et le troisième graphique, les p-valeurs individuelles ainsi que celle du couple sont beaucoup plus faibles que dans la figure 2.4.

On peut voir à la figure 2.6 que les marqueurs détectés pour le second échantillon sont plus près des vrais TIMs que dans le premier dans l'ensemble. Par contre, on est incapable d'en détecter un des deux de façon exacte contrairement à l'exemple précédent. On voit dans le nuage de points que les marqueurs près des deux lignes rouges sont légèrement dépassés par des points les avoisinant. Ceci s'illustre également dans le graphique du chi-carré pour des couples de marqueurs et dans celui de la logistique également. Les détections par ces deux dernières méthodes sont identiques ou presque. Pour ce qui est de la figure 2.7, on peut remarquer qu'une fois de plus, les TIMs réels sont détectés de façon indubitable par les trois méthodes lorsqu'ils sont présents dans les données.

Dans le troisième échantillon, si on observe la figure 2.8, la détection du premier marqueur causal est bonne. On voit dans le graphique du haut que les marqueurs à sa gauche semblent plus associés à maladie que ceux à sa droite et que c'est le marqueur le plus près de lui dans ce premier groupe qui a la plus petite p-valeur. C'est celui qu'on semble détecter comme première composante du couple dans les deuxième et troisième graphiques également. On ne voit aucune association significative pour ce qui est du second TIM dans aucun des trois graphiques, ce à quoi on peut s'attendre par sa très faible contribution à l'incidence de la maladie

si on le compare au premier. En fait, le second marqueur qui semble être détecté par le khi-carré multiple comme par la logistique est le premier marqueur, mais une deuxième fois (les deux détections sont près de 0.45). C'est également le seul échantillon où rajouter les vraies mutations causales dans les données ne semble pas garantir une détection parfaite. Par contre, les p-valeurs des couples détectés diminuent encore une fois dans la figure 2.9 par rapport à la figure 2.8.

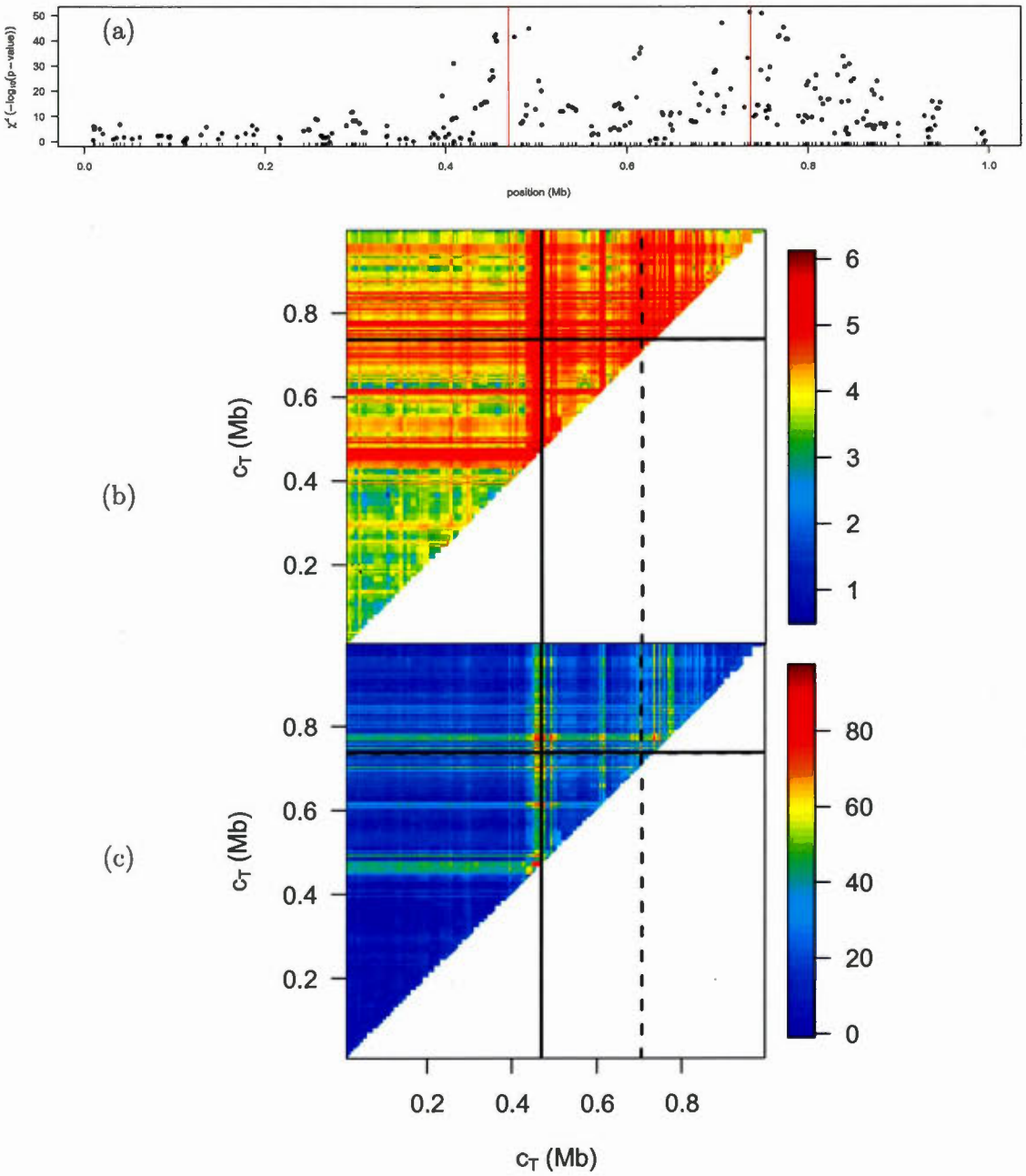

Figure 2.4: Premier modèle sans les TIMs. L'inverse du logarithme des p-valeurs pour le χ^2 (a) du test unidimensionnel, (b) du test bidimensionnel et (c) approximant le ratio de log-vraisemblance pour le modèle à 2 gènes. 500 cas et 500 témoins.

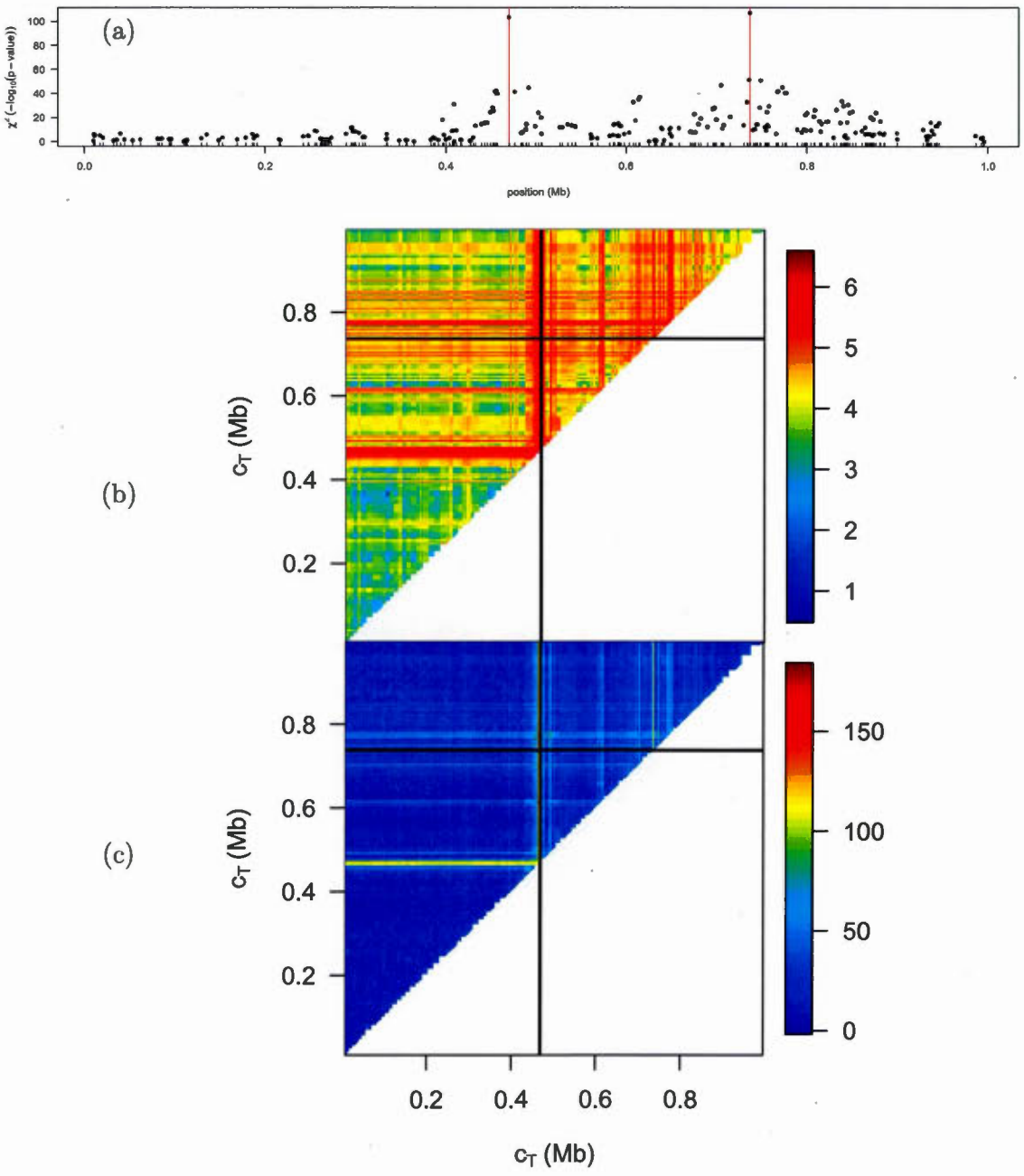

Figure 2.5: Premier modèle avec les TIMs. L'inverse du logarithme des p-valeurs pour le χ^2 (a) du test unidimensionnel, (b) du test bidimensionnel et (c) approximant le ratio de log-vraisemblance pour le modèle à 2 gènes. 500 cas et 500 témoins.

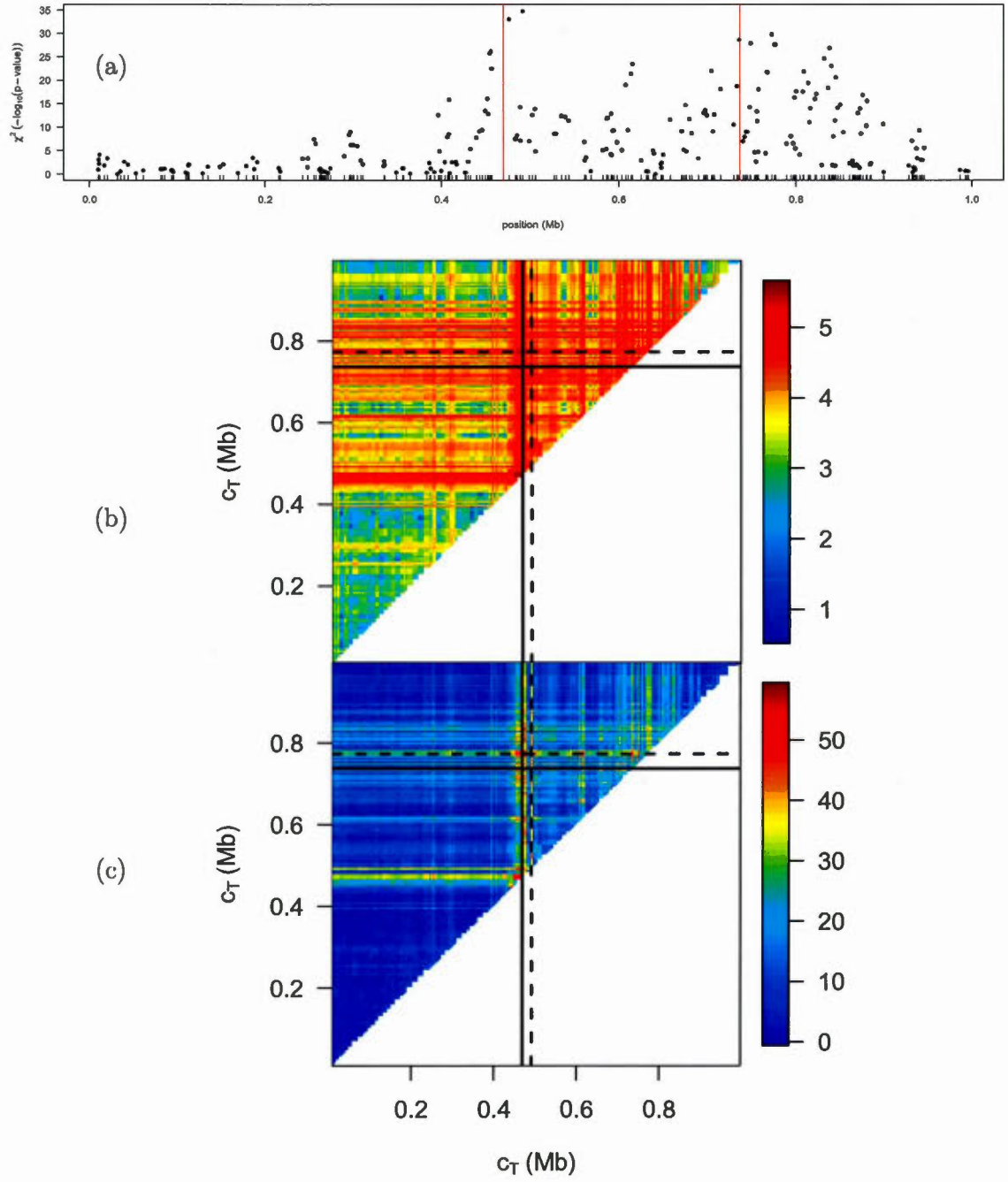

Figure 2.6: Deuxième modèle sans les TIMs. L'inverse du logarithme des p-valeurs pour le χ^2 (a) du test unidimensionnel, (b) du test bidimensionnel et (c) approximant le ratio de log-vraisemblance pour le modèle à 2 gènes. 400 cas et 400 témoins.

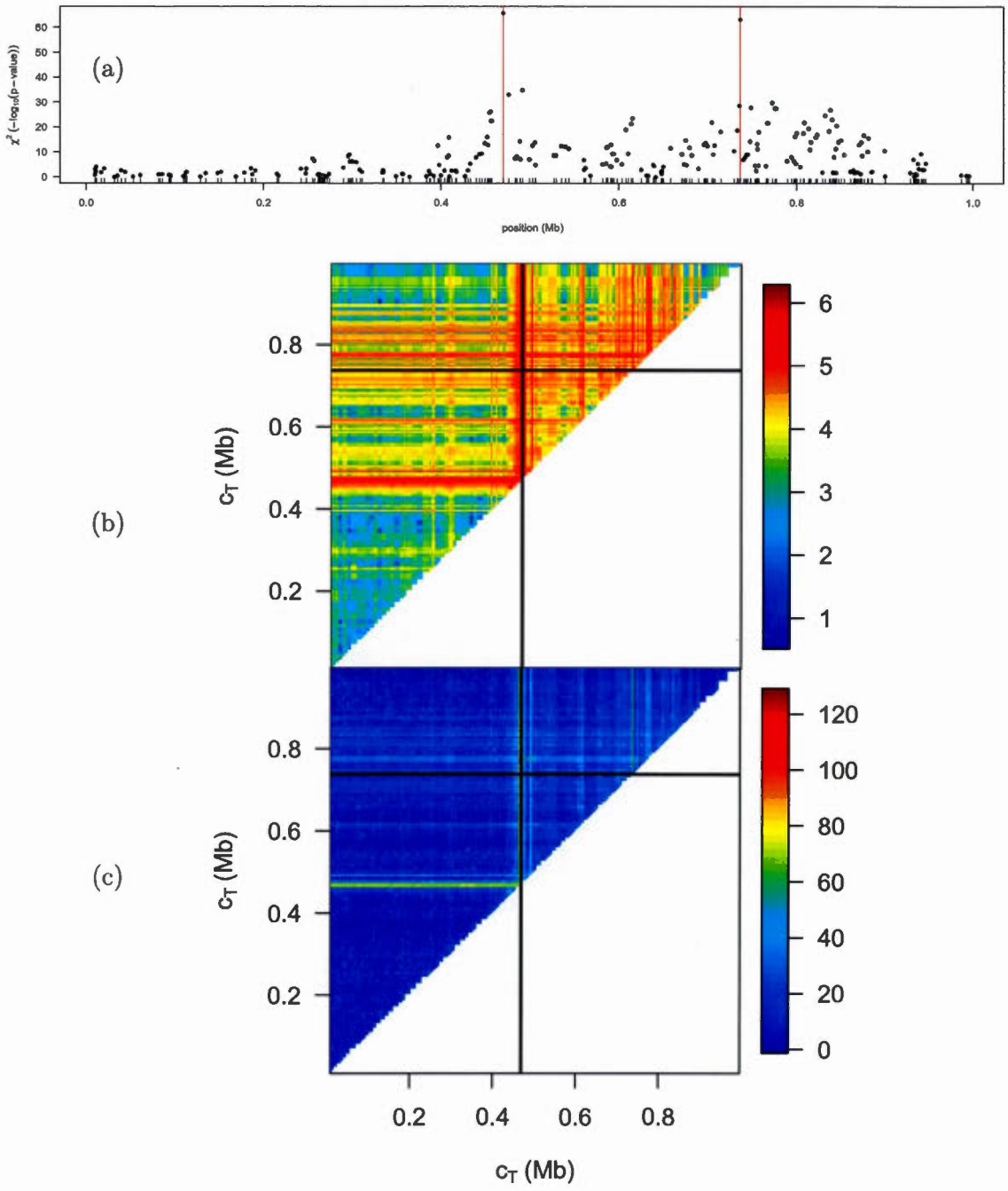

Figure 2.7: Deuxième modèle avec les TIMs. L'inverse du logarithme des p-valeurs pour le χ^2 (a) du test unidimensionnel, (b) du test bidimensionnel et (c) approximant le ratio de log-vraisemblance pour le modèle à 2 gènes. 400 cas et 400 témoins.

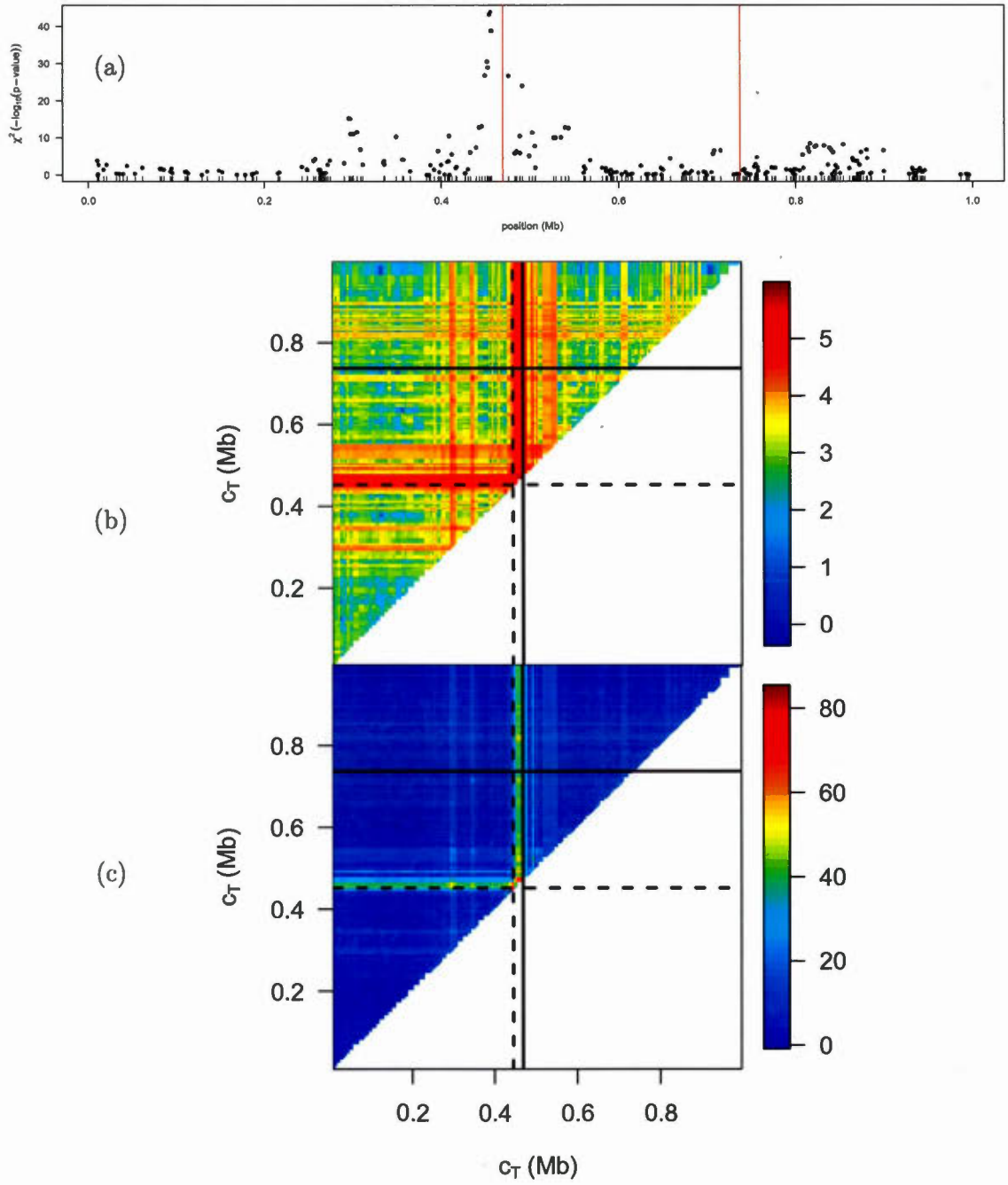

Figure 2.8: Troisième modèle sans les TIMs. L'inverse du logarithme des p-valeurs pour le χ^2 (a) du test unidimensionnel, (b) du test bidimensionnel et (c) approximant le ratio de log-vraisemblance pour le modèle à 2 gènes. 400 cas et 400 témoins.

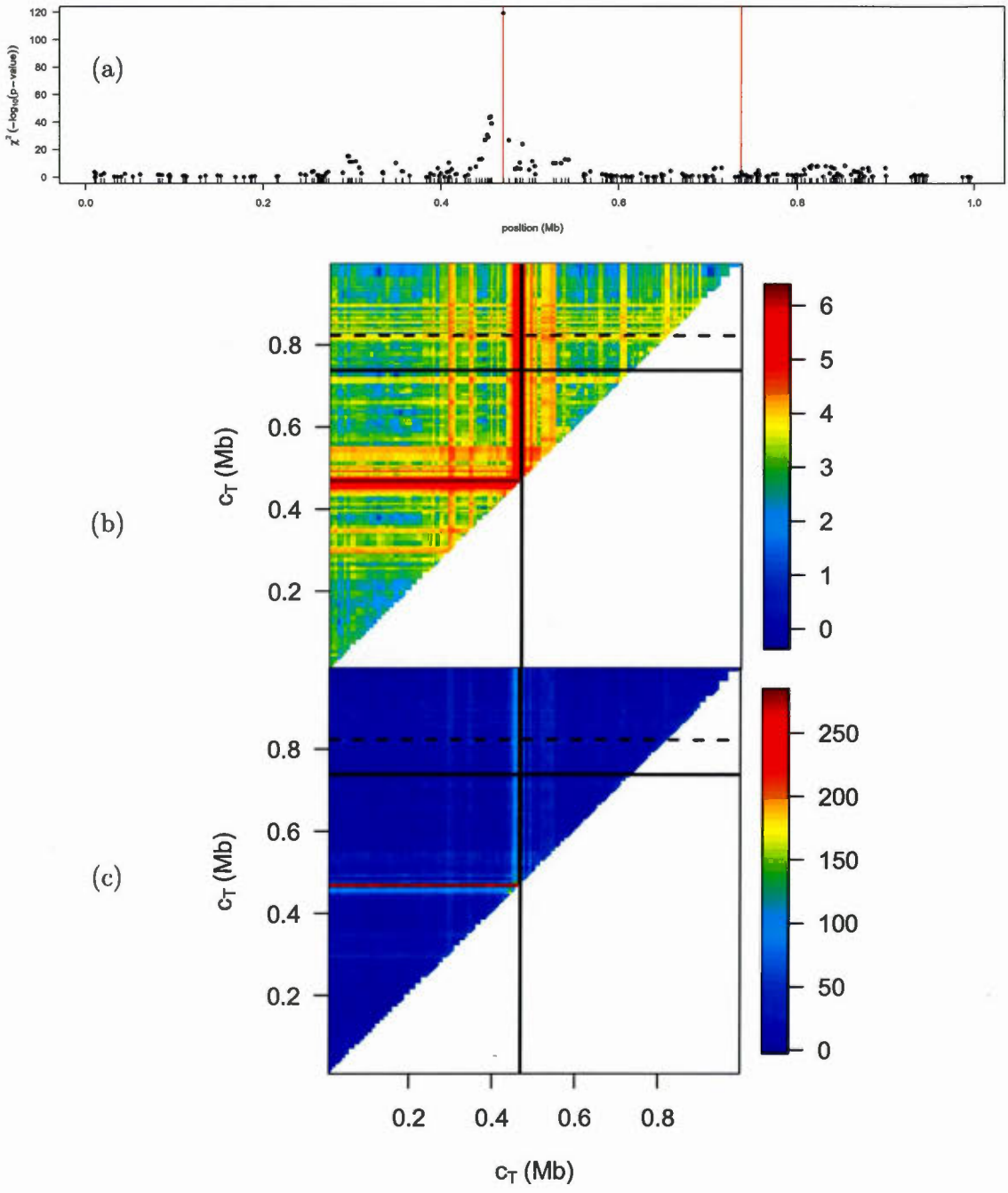

Figure 2.9: Troisième modèle avec les TIMs. L'inverse du logarithme des p-valeurs pour le χ^2 (a) du test unidimensionnel, (b) du test bidimensionnel et (c) approximant le ratio de log-vraisemblance pour le modèle à 2 gènes. 400 cas et 400 témoins.

CHAPITRE III

LE PROCESSUS DE COALESCENCE

Le processus de coalescence, un modèle publié originellement dans l'article *The Coalescent* de Kingman (1982), est un processus stochastique très important en génétique. Il sera montré dans ce chapitre que, bien que basé initialement sur des hypothèses simples, il est possible de le modifier afin de l'utiliser à bien des fins. C'est la base théorique de la méthodologie sur laquelle ce mémoire porte et un outil puissant de simulation de généalogies. On verra dans ce chapitre, tout d'abord, le modèle duquel lui-même s'inspire. Puis, le processus de coalescence classique sera présenté pour enfin être complexifié en y incluant des phénomènes qui causent une diversité génétique importante chez l'être humain.

3.1 Le modèle de Wright-Fisher

Le modèle neutre de Wright-Fisher, qui fut développé par Fisher (1922) puis Wright (1931), est un modèle stochastique de reproduction de la population. Il est basé sur des hypotèses relativement simples, qui lui ont permis d'être un outil profitable en génétique des populations, car les complexifications qu'on peut lui apporter sont nombreuses.

Le cadre dans lequel on se trouve est celui d'une population finie de N individus diploïdes, et donc $2N$ individus haploïdes (dans ce mémoire en particulier, ces

individus correspondent à des séquences génétiques, qui représentent des sections de chromosome relativement courtes). Les générations sont discrètes et distinctes (toutes les séquences d'une génération naissent et meurent au même moment) et leur taille est invariable.

Ainsi, nous avons une première génération de $2N$ individus qui ont existé dans un passé plus ou moins lointain. La prochaine génération consistera, pour sa part, en un échantillon aléatoire avec remise de $2N$ individus tirés de la population originelle. Comme il n'y a pas de sélection dans le modèle neutre de Wright-Fisher, les probabilités qu'ont chacun des individus d'être transmis une seule fois à la génération prochaine est $1/2N$. Et l'expérience se répète jusqu'à la génération actuelle. Comme les échantillons sont tirés avec remise, les expériences sont indépendantes. Ainsi, la probabilité que l'individu i de la génération k soit transmis j fois à la génération $k+1$ est issue d'une loi binomiale de paramètres $n = 2N$ et $p = 1/2N$.

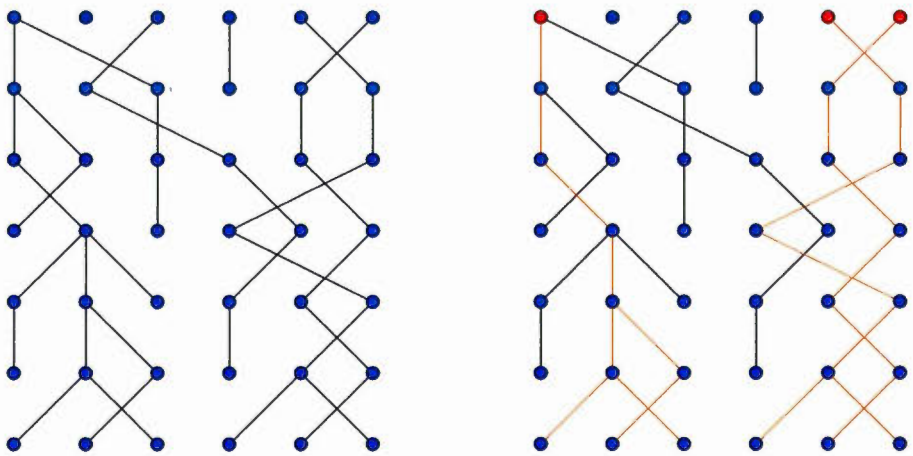

Figure 3.1: Exemple d'application du modèle de Wright-Fisher pour 7 générations de $2N = 6$ individus (gauche) et visualisation des lignées encore vivantes aujourd'hui (droite).

La figure 3.1 est un exemple de simulation d'une généalogie comportant 7 générations pour une population de 6 individus selon le modèle de Wright-Fisher. On peut remarquer que plusieurs des lignées se sont éteintes avant la présente génération, les 3 points rouges de l'image de droite représentant les seules séquences ayant au moins un descendant dans la génération la plus récente (les lignées oranges correspondent aux généalogies des membres de la population faisant partie de cette génération). Ainsi, simuler de cette façon, du passé vers le présent, n'est pas la façon la plus efficace de faire, étant donné la quantité de travail faite pour simuler des histoires généalogiques dont les descendants ne seront jamais observés. La section qui suit démontre une solution qui a été trouvée il y a quelques décennies par un mathématicien britannique afin d'éviter ce désagrément et d'optimiser la génération de généalogies.

3.2 Le processus de coalescence

Kingman (1982) est celui qui introduit le processus de coalescence. Ce dernier est un processus de Markov dont l'espace d'états est discret et qui est basé sur le modèle de Wright-Fisher. Bien que le temps soit continu dans ce processus, l'idée qui a mené à celui-ci était basée sur un processus à temps discret.

Supposons que l'on a $2N$ séquences observables dans le présent dans notre population. Le processus de coalescence à temps discret consiste à déterminer quelles sont les séquences contenues dans les générations qui ont précédé celle-ci dans l'histoire en remontant dans le passé, une génération à la fois. On ne considère dans notre espace d'états à une génération que les séquences ayant pu donner naissance aux séquences présentes dans la génération qui le suit, à la différence du modèle de Wright-Fisher qui décrit même les lignées éteintes dans le présent. Le terme coalescence en lui-même, en génétique, représente la fusion de deux ou plusieurs lignées ou, en d'autres mots, lorsque deux ou plusieurs enfants trouvent

un même parent. La probabilité que deux séquences i et j présentes dans la génération actuelle coalescent une génération dans le passé est

$$P(C_{ij}) = \frac{1}{2N}. \tag{3.1}$$

Donc, la probabilité que k générations s'écoulent avant que la coalescence entre i et j ait lieu, l'évènement noté C_{ij}^{k}, découle d'une loi géométrique de paramètre $p = 1/2N$ et est ainsi celle énoncée dans (3.2), car on veut qu'il y ait $k-1$ transitions où les lignées de i et j ne coalescent pas suivies d'un évènement de coalescence entre les deux :

$$P(C_{ij}^{k}) = \left(1 - \frac{1}{2N}\right)^{k-1} \frac{1}{2N}. \tag{3.2}$$

Maintenant, généralisons ce raisonnement à un contexte où nous considérons un échantillon de n séquences d'une certaine génération de notre histoire généalogique. La probabilité que les n lignées demeurent distinctes lors du passage à la génération précédente, $P(n)$, exprimée à l'équation (3.3), est celle que les n séquences choisissent des ancêtres distincts parmi les $2N$ possibles :

$$\begin{aligned}
P(n) &= \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)\ldots\left(\frac{2N-(n-1)}{2N}\right) && (3.3)\\
&= \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\ldots\left(1-\frac{n-1}{2N}\right)\\
&= 1-\frac{1}{2N}-\frac{2}{2N}\ldots-\frac{n-1}{2N}+\mathcal{O}\left(\frac{1}{N^2}\right)\\
&= 1-\frac{\sum_{k=1}^{n-1}k}{2N}+\mathcal{O}\left(\frac{1}{N^2}\right)\\
&= 1-\frac{\sum_{k=1}^{n-1}k}{2N}+\mathcal{O}\left(\frac{1}{N^2}\right)\\
&= 1-\frac{n(n-1)/2}{2N}+\mathcal{O}\left(\frac{1}{N^2}\right)\\
&= 1-\frac{\binom{n}{2}}{2N}+\mathcal{O}\left(\frac{1}{N^2}\right) && (3.4)
\end{aligned}$$

$$\approx 1 - \frac{\binom{n}{2}}{2N}. \tag{3.5}$$

Comme le terme $\mathcal{O}(1/N^2)$ devient négligeable lorsque notre population est très grande et qu'il représente la probabilité des coalescences de plus de deux lignées à la fois, il est acceptable de ne pas considérer les évènements de ce type dans le modèle, d'où le passage de (3.4) à (3.5). Cela nous permet donc d'approximer la probabilité qu'il n'y ait aucun évènement de coalescence lors du passage d'une génération de n individus à la génération précédente par (3.5).

Ainsi, la probabilité que cela prenne exactement k générations avant qu'une première coalescence ait lieu dans un échantillon de n séquences, T_c^n, faisant passer sa taille à $n-1$, est donnée par une loi géométrique de paramètre $1-\left(1-\binom{n}{2}/2N\right) = \binom{n}{2}/2N$:

$$P(T_c^n = k) = \left(1 - \frac{\binom{n}{2}}{2N}\right)^{k-1} \frac{\binom{n}{2}}{2N}. \tag{3.6}$$

Étant donné que le temps moyen en terme de nombre de générations avant qu'il y ait coalescence,

$$E(T_c^n) = \frac{1}{\frac{\binom{n}{2}}{2N}} = \frac{2N}{\binom{n}{2}}, \tag{3.7}$$

devient très grand quand n décroît, on procède à un changement d'échelle où on considère que $2N$, qui représente le temps moyen de coalescence entre 2 séquences devient l'unité de temps de référence. En considérant maintenant que notre $2N$ tend vers l'infini, il est possible d'approximer la loi géométrique par une loi exponentielle de paramètre $\lambda = \binom{n}{2}$, ce qui est fait dans le processus de coalescence classique (souvenons-nous que nous avons dit au début de la section que le processus de coalescence était une approximation du modèle de Wright-Fisher). Ce choix est avantageux du point de vue de l'efficacité, car si on génère des généalogies avec le processus de coalescence à temps discret, on obtient plusieurs générations où il

n'y a aucune coalescence, surtout quand n diminue, ce qui amène la réalisation de beaucoup d'opérations sans réel intérêt. Or, avec le processus de coalescence classique, il ne suffit que de générer une suite de temps exponentiels qui nous indiquent quels sont les intervalles de temps écoulés entre les coalescences qui ont mené au MRCA, de l'anglais *Most Recent Common Ancestor*, le dernier ancêtre partagé par nos n séquences.

3.3 Le processus de coalescence avec mutations

Comme indiqué précédemment dans le chapitre, l'avantage de modèles basés sur très peu d'hypothèses, comme le modèle de Wright-Fisher et son approximation, le processus de coalescence, est que ça permet de les complexifier plus aisément, en tenant compte des besoins de chacun.

On a vu dans le chapitre 1 qu'un enfant héritait très rarement du bagage génétique inchangé de ses parents et qu'un des phénomènes biologiques responsables de cette diversité génétique est la mutation génétique. Le type de mutations qui nous intéresse dans le cadre de ce mémoire est le SNP respectant le modèle des sites infinis, un modèle introduit par Watterson (1975). L'idée sur laquelle repose celui-ci est qu'étant donné que notre ADN est extrêmement long, si on suppose que le nombre de sites contenus par celui-ci est infini et que les évènements de mutation dont on traite sont assez rares, la probabilité qu'une mutation arrive deux fois sur le même site dans l'histoire de la généalogie est négligeable («la foudre ne frappe jamais deux fois au même endroit», dit-on). Ainsi, lorsqu'une séquence subit une mutation respectant le modèle des sites infinis, elle la transmet à tous ses descendants sans exception.

Supposons qu'une séquence génétique devient porteuse d'une mutation génétique selon un taux de μ par génération (on peut faire varier ce taux selon la longueur de notre séquence). Comme la probabilité d'une mutation est de l'ordre de 10^{-10},

on considère que l'apparition de deux mutations ou plus durant une génération est hautement improbable. La probabilité que le temps avant de subir une mutation, T_M, soit de k générations suit une loi géométrique de paramètre $p = \mu$:

$$P(T_M = k) = (1 - \mu)^{k-1}\mu. \tag{3.8}$$

Comme pour le processus de coalescence classique, on approxime la loi géométrique décrivant T_M par une exponentielle en utilisant le principe que la loi géométrique tend asymptotiquement vers une exponentielle lorsque $2N$ tend vers l'infini et en réécrivant μ comme $2N\mu/2N$. Dans ce cas-ci, $\lambda = 2N\mu$, $2N$ étant encore une fois notre unité de temps, pour que l'échelle soit la même que dans la section précédente. En appliquant la transformation utilisée par souci de simplification dans la littérature $\theta = 4N\mu$, le paramètre rephasé de la loi exponentielle est $\lambda = \theta/2$.

Donc, le temps minimum avant qu'il y ait une mutation dans un échantillon de n lignées est le minimum de n lois exponentielles indépendantes, toutes de paramètre $\lambda = \theta/2$, ce qui en fait un temps exponentiel de paramètre $\lambda = n\theta/2$.

En considérant le processus de coalescence avec mutations, à chaque moment de notre histoire généalogique où on considère qu'il y a n séquences, le prochain évènement à arriver lorsqu'on recule dans le temps peut être soit une coalescence entre deux séquences, qui va réduire la taille de notre échantillon de séquences de 1, ou une mutation sur une des séquences, qui va laisser la taille de notre échantillon inchangée. Comme le temps avant un évènement de coalescence suit une loi exponentielle de paramètre $\lambda = \binom{n}{2}$, le temps avant un évènement de mutation suit pour sa part une loi exponentielle de paramètre $\lambda = n\theta/2$ et les évènements des deux types sont indépendants, le temps avant qu'un évènement

arrive, peu importe sa catégorie, suit une loi exponentielle de paramètre :

$$\begin{aligned} \lambda &= \binom{n}{2} + \frac{n\theta}{2}, \\ &= \frac{n(n-1)}{2} + \frac{n\theta}{2}, \\ &= \frac{n(n-1+\theta)}{2}. \end{aligned} \tag{3.9}$$

Cela fait en sorte que le temps moyen avant un évènement, soit de coalescence ou de mutation, est $2/n(n-1+\theta)$ unités de $2N$ générations. Une fois le temps simulé, il faut être en mesure de savoir lequel des évènements se produit. En usant des propriétés des lois exponentielles, la probabilité que cet évènement soit une coalescence, C, ou une mutation, M, sont respectivement :

$$P(C) = \frac{\frac{n(n-1)}{2}}{\frac{n(n-1+\theta)}{2}} = \frac{n-1}{n-1+\theta}, \tag{3.10}$$

$$P(M) = \frac{\frac{n\theta}{2}}{\frac{n(n-1+\theta)}{2}} = \frac{\theta}{n-1+\theta}. \tag{3.11}$$

3.4 Le graphe de recombinaison ancestral

La recombinaison a été introduite dans le chapitre 1 comme autre source majeure de diversité génétique chez l'être humain. C'est Hudson (1983) qui introduit ce phénomène d'une importance majeure au processus de coalescence. On se souvient que, lorsqu'observée dans l'ordre naturel des choses, la recombinaison a lieu lorsque deux chromosomes formant une paire échangent une partie de leur matériel génétique durant la méiose, le chromosome étant transmis à l'enfant par son parent étant donc un hybride entre le chromosome hérité de son grand-père et celui hérité de sa grand-mère.

Ainsi, pour reconstruire ce phénomène du présent vers le passé, on peut se dire que le parent d'une séquence correspond à deux séquences, une ayant contribué au matériel à gauche d'un point de recombinaison et l'autre au matériel à droite

de ce même point. On complète la première séquence avec du matériel dont on ne connaît la composition du point de recombinaison jusqu'à son extrémité droite, puis on fait l'inverse pour la deuxième séquence, car la longueur de nos séquences est invariable à travers le temps dans le processus de coalescence. Ce matériel utilisé pour la complétion des séquences est nommé matériel non ancestral et n'a pas de réelle utilité dans le modèle, car il n'a pas été transmis à notre échantillon le plus récent. Le point de recombinaison est distribué de façon uniforme sur l'entièreté de la séquence. Une représentation graphique des deux perceptions de la recombinaison se trouve à la figure 3.2 (les quatre différentes couleurs représentent différentes bases azotées A,C,T et G). On voit dans l'image de gauche que le point de recombinaison est quelque part entre le 3^e et le 4^e marqueur génétique de notre séquence, ce qui fait en sorte qu'elle hérite les trois premiers marqueurs du premier parent et les deux derniers du deuxième parent. En partant de l'enfant et en connaissant la position du point de recombinaison, on peut réattribuer les trois marqueurs de gauche à un parent et les deux de droite à l'autre (et compléter avec du matériel non ancestral, représenté par les marqueurs gris).

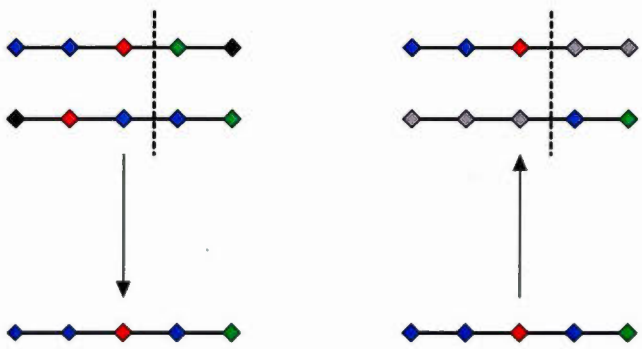

Figure 3.2: Recombinaison vue du passé vers le présent (gauche) et vue du présent vers le passé (droite). En gris, le matériel non ancestral.

Étant donné que maintenant une séquence peut avoir plus d'un parent, nous

n'avons plus affaire à un arbre comme dans les deux dernières sections, mais à un graphe, ce qui explique que le modèle de Hudson porte le nom de graphe de recombinaison ancestral, souvent abrégé par le terme ARG en raison de l'anglais *Ancestral Recombination Graph*. Par contre, si on prend la position d'un marqueur donné sur notre séquence, on est encore en mesure de lui associer un arbre afin d'illustrer sa généalogie, il s'agit de choisir la branche allant vers la séquence ayant transmis ce marqueur lorsqu'il y a un évènement de recombinaison.

Considérons maintenant qu'il y a recombinaison sur une séquence selon un taux de r par génération (on peut également faire varier ce taux selon la longueur des séquences de notre simulation). La probabilité que le temps avant de subir une recombinaison, T_R, soit de k générations suit une loi géométrique de paramètre $p = r$:

$$P(T_R = k) = (1 - r)^{k-1} r. \tag{3.12}$$

On approxime encore une fois la loi géométrique décrivant T_R par une exponentielle de taux $\lambda = 2Nr$ et on utilise une transformation analogue à celle utilisée pour les mutations, $\rho = 4Nr$. Donc, le paramètre transformé est $\lambda = \rho/2$. Le temps minimum avant qu'il y ait une mutation dans un échantillon de n lignées est ainsi le minimum de n lois exponentielles indépendantes, toutes de paramètre $\lambda = \rho/2$, ce qui en fait un temps exponentiel de paramètre $\lambda = n\rho/2$.

Comme les recombinaisons ont lieu indépendamment des coalescences et des mutations, le temps avant qu'il y ait un premier évènement appartenant à un de ces 3 types suit une loi exponentielle de paramètre :

$$\begin{aligned} \lambda &= \binom{n}{2} + \frac{n\theta}{2} + \frac{n\rho}{2}, \\ &= \frac{n(n-1)}{2} + \frac{n\theta}{2} + \frac{n\rho}{2}, \end{aligned}$$

$$= \frac{n(n-1+\theta+\rho)}{2}. \tag{3.13}$$

Le temps moyen avant un prochain évènement soit de coalescence, de mutation ou de recombinaison est maintenant $2/n(n-1+\theta+\rho)$ en unités de $2N$ générations. Les probabilités que cet évènement soit une coalescence, C, une mutation, M, ou une recombinaison, R, sont respectivement :

$$P(C) = \frac{\frac{n(n-1)}{2}}{\frac{n(n-1+\theta+\rho)}{2}} = \frac{n-1}{n-1+\theta+\rho}, \tag{3.14}$$

$$P(M) = \frac{\frac{n\theta}{2}}{\frac{n(n-1+\theta+\rho)}{2}} = \frac{\theta}{n-1+\theta+\rho}, \tag{3.15}$$

$$P(R) = \frac{\frac{n\rho}{2}}{\frac{n(n-1+\theta+\rho)}{2}} = \frac{\rho}{n-1+\theta+\rho}. \tag{3.16}$$

On peut se poser la question quant à la convergence de notre graphe de recombinaison ancestral vers un échantillon de taille $n = 1$, le MRCA, car les recombinaisons augmentent la taille de notre échantillon. La réponse est qu'elle existe toujours, car le taux des évènements de coalescence, qui eux réduisent la taille de l'échantillon, est quadratique en n ($\lambda = n(n-1)/2$) alors que le taux des évènements de recombinaison est linéaire en n ($\lambda = n\rho/2$).

En définitive, l'algorithme complet pour générer un graphe de recombinaison ancestral avec mutations en débutant avec un échantillon de n séquences s'effectue comme suit :

1. Poser $k = n$.
2. Si $k = 1$, terminer l'algorithme.

   Si $k > 1$, simuler un temps exponentiel de taux $\lambda = n(n-1+\theta+\rho)/2$.
3. Déterminer si le prochain évènement est une coalescence avec probabilité $(n-1)/(n-1+\theta+\rho)$, une mutation avec probabilité $\theta/(n-1+\theta+\rho)$ ou une recombinaison avec probabilité $\rho/(n-1+\theta+\rho)$.

4. Une fois le type d'évènement sélectionné,
   (a) si celui-ci est une coalescence, choisir aléatoirement deux séquences qui doivent avoir le même matériel ancestral et les faire coalescer. Poser $k = k - 1$.
   (b) si celui-ci est une mutation, choisir aléatoirement une séquence et enlever une mutation d'un de ces loci ancestraux mutants (étant donné qu'on remonte dans le temps).
   (c) si celui-ci est une recombinaison, choisir aléatoirement une séquence puis un point de recombinaison sur cette séquence. La diviser en une séquence gauche et une séquence droite, et compléter celles-ci avec du matériel inconnu. Poser $k = k + 1$.
5. Retourner à l'étape 2.

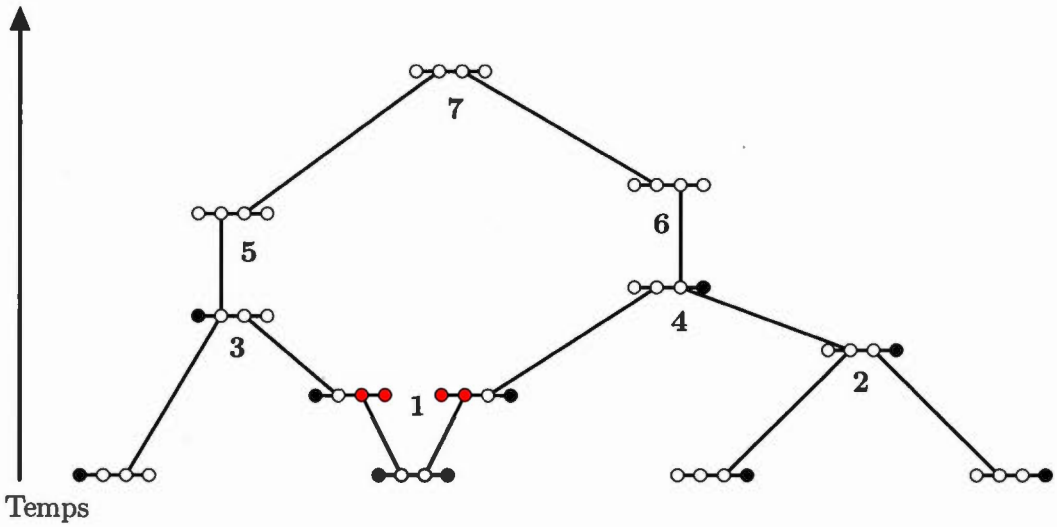

Figure 3.3: ARG de $n = 4$ séquences qui trouvent leur plus récent ancêtre commun en 7 étapes.

Nous pouvons observer à la figure 3.3 un exemple d'ARG dont l'échantillon de

départ contient $n = 4$ séquences où le matériel primitif est représenté en blanc, le matériel provenant d'une mutation dans l'histoire de la généalogie en noir et le matériel n'ayant été transmis au dit échantillon de départ (ou matériel non ancestral) en rouge. Ces séquences convergent vers leur MRCA après 7 évènements, numérotés dans la figure :

1. Le premier évènement à arriver lorsqu'on remonte dans le temps à partir de notre séquence observée jusqu'au moment où le MRCA est trouvé est une recombinaison de la deuxième séquence de l'échantillon où le chiasmata est situé entre le deuxième et le troisième marqueur. Les séquences résultantes sont donc complétées avec du matériel non ancestral. Le nombre de séquences dans cette génération est $k = n + 1 = 5$.

2. Le deuxième évènement est une coalescence entre les deux dernières séquences de l'échantillon observé, qui ont exactement la même séquence génétique. Le nombre de séquences dans cette génération est $k = k - 1 = 4$.

3. Le troisième évènement est une coalescence entre la première séquence de l'échantillon et la séquence résultant de la recombinaison de l'étape 1 contenant le matériel ancestral à gauche du chiasmata. Il est possible pour ces deux séquences de coalescer, car leur matériel ancestral (qui n'est pas rouge) est identique. Nous connaissons donc maintenant la partie manquante de la séquence contenant du matériel non ancestral. Le nombre de séquences dans cette génération est $k = k - 1 = 3$.

4. Le quatrième évènement est une coalescence entre la séquence résultant du deuxième évènement et la séquence résultant de la recombinaison de l'étape 1 contenant le matériel ancestral à droite du chiasmata. Elles ont le même matériel ancestral. Le nombre de séquences dans cette génération est $k = k - 1 = 2$.

5. Le cinquième évènement est la perte d'une mutation au premier marqueur

de la séquence résultant de l'étape 3. Le nombre de séquences dans cette génération est $k = 2$.

6. Le sixième évènement est la perte d'une mutation au dernier marqueur de la séquence résultant de l'étape 4. Le nombre de séquences dans cette génération est $k = 2$.

7. Le septième et dernier évènement est une coalescence entre les séquences résultant des étapes 5 et 6, qui sont toutes deux des séquences n'ayant que du matériel primitif. Le nombre de séquences de cette génération est $k = k - 1 = 1$. Nous avons obtenu le MRCA de nos 4 séquences.

CHAPITRE IV

DMAPINTERACTION : L'EXTENSION DE *DMAP*

Une méthodologie visant à utiliser le processus de coalescence afin de faire de la cartographie génétique fine dans le cas de maladies causées par une mutation de type SNP fut développée par Descary (2012). Cette méthodologie, implémentée en C++, porte le nom de *DMap*. Mais que se passe-t'il si on traite d'une maladie causée par l'interaction de deux SNPs comme on a vu qu'il en existe ? C'est la question à laquelle nous tenterons de répondre dans ce chapitre.

4.1 Une introduction à *DMap* classique

C'est un échantillon de n individus et ainsi $2n$ séquences qui est utilisé par *DMap* afin de faire de la cartographie génétique. Dénotons cet ensemble par H_0, qui est la génération actuelle de séquences de notre généalogie. *DMap* peut construire un nombre variable de graphes de recombinaison ancestraux (ARGs) qui ont une caractéristique commune : leur point de départ lorsqu'on remonte du présent vers le passé est H_0. Ces ARGs représentent donc tous des généalogies qui auraient pu s'avérer être l'histoire génétique de notre échantillon. Certaines de ces généalogies sont néanmoins plus vraisemblables que d'autres, d'où le besoin de déterminer la probabilité de chacune. En considérant que la probabilité d'une généalogie G est celle d'observer sa suite de $\tau^* + 1$ états notés $H_0, H_1, ..., H_{\tau^*}$ où H_{τ^*} est le MRCA, nous utilisons la propriété markovienne du processus de coalescence (chaque état

ne dépend que de son prédécesseur direct) pour la déterminer :

$$\begin{aligned} P(G) &= P(H_0, H_1, H_2, ..., H_{\tau^*}) \\ &= P(H_0|H_1, H_2, ..., H_{\tau^*})P(H_1, H_2, ..., H_{\tau^*}) \\ &= P(H_0|H_1)P(H_1, H_2, ..., H_{\tau^*}) \\ &= P(H_0|H_1)P(H_1|H_2)P(H_2, ..., H_{\tau^*}) \\ &= ... \\ &= P(H_{\tau^*}) \prod_{\tau=0}^{\tau^*-1} P(H_\tau|H_{\tau+1}) \qquad (4.1) \\ &= \prod_{\tau=0}^{\tau^*-1} P(H_\tau|H_{\tau+1}). \qquad (4.2) \end{aligned}$$

Le passage de l'équation (4.1) à l'équation (4.2) s'explique par la connaissance que nous avons de H_{τ^*} : il ne contient que notre MRCA, qui s'exprime comme une suite de marqueurs ancestraux. Donc, $P(H_{\tau^*}) = 1$.

Dans notre contexte, un graphe est une suite d'évènements de type coalescence, mutation ou recombinaison qui nous mèneront de chaque état τ à l'état $\tau + 1$ pour $\tau \in \{0, 1, ..., \tau^* - 1\}$. Nous avons vu au chapitre précédent que les probabilités théoriques que le prochain évènement dans notre processus soit une coalescence, une mutation ou une recombinaison lorsque l'on a n séquences sont respectivement

$$\frac{n-1}{n-1+\theta+\rho}, \qquad (4.3)$$

$$\frac{\theta}{n-1+\theta+\rho}, \qquad (4.4)$$

$$\frac{\rho}{n-1+\theta+\rho}. \qquad (4.5)$$

Or, étant donné que les évènements de recombinaison ont comme conséquence qu'une partie du matériel génétique présent dans les séquences est non ancestral et que, étant donné que ce matériel n'est pas transmis à H_0, il n'a pas d'intérêt

pour nous, nous devons modifier nos probabilités pour l'omettre. Considérons que nos n séquences à chaque génération τ peuvent être subdivisées en k types de séquences distinctes. Considérons également que chaque type i de séquence se retrouve n_i fois dans H_τ, $i \in \{1, ..., k\}$ de telle sorte que $\sum_{i=1}^{k} n_i = n$. De plus, définissons a_i comme étant le nombre de marqueurs ancestraux contenus dans une séquence de type i, r_i comme étant la distance entre les deux marqueurs ancestraux les plus éloignés dans une même séquence, p comme le nombre de marqueurs (ancestraux ou non) par séquence et r comme la longueur totale d'une séquence. On peut donc calculer α, la proportion de marqueurs contenus dans nos séquences qui sont ancestraux, et β, la proportion de la longueur de nos séquences où une recombinaison qui nous intéresse est possible, comme dans les équations (4.6) et (4.7) :

$$\alpha = \sum_{i=1}^{k} \frac{n_i a_i}{np}, \tag{4.6}$$

$$\beta = \sum_{i=1}^{k} \frac{n_i r_i}{nr}. \tag{4.7}$$

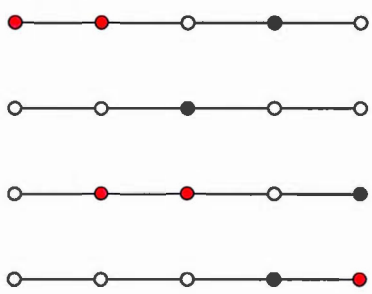

Figure 4.1: Génération H_τ contenant $n = 4$ séquences contenant $p = 5$ marqueurs et de longueur totale $r = 8$ Mb chacune.

L'exemple d'une génération H_τ contenant $n = 4$ séquences tirée d'une généalogie est illustré dans la figure 4.1. Elles sont toutes de longueur $r = 8$ Mb et on

Type i de séquence	n_i	a_i	r_i (Mb)
1	1	3	4
2	1	5	8
3	1	3	8
4	1	4	8

Tableau 4.1: Paramètres nécessaires au calcul de α et β de chacune des quatre séquences illustrées de haut en bas à la figure 4.1.

considère que les 5 marqueurs sont équidistants, il y a donc une distance de 2 Mb entre chacun. Les marqueurs ancestraux sont blancs s'ils sont primitifs ou noirs s'ils sont mutants alors que les marqueurs non ancestraux sont rouges. Les divers paramètres nécessaires pour calculer α et β pour chacune des séquences sont indiqués dans le tableau 4.1. Ainsi, dans cet échantillon, on peut déterminer :

$$\begin{aligned} \alpha &= \frac{1\cdot 3+1\cdot 5+1\cdot 3+1\cdot 4}{4\cdot 5} = \frac{3}{4}, \\ \beta &= \frac{1\cdot 4+1\cdot 8+1\cdot 8+1\cdot 8}{4\cdot 8} = \frac{7}{8}. \end{aligned}$$

Comme nous ne nous soucions que des mutations ayant lieu sur les marqueurs qui sont du matériel ancestral, le taux auxquel elles se produisent n'est plus θ, mais bien $\alpha\theta$. Également, comme les recombinaisons ayant lieu sur des extrémités non ancestrales de nos séquences n'ont pas d'intérêt pour nous, car on se retrouverait suite à une recombinaison de ce type avec une séquence ne contenant que du matériel n'ayant pas été transmis à nos données de départ, le taux de recombinaison n'est plus ρ, mais $\beta\rho$.

Les probabilités que le prochain évènement dans notre processus soit une coalescence C, une mutation M ou une recombinaison R lorsque l'on a n séquences

contenues dans un H_τ donné sont donc maintenant :

$$P(C) = \frac{n-1}{n-1+\alpha\theta+\beta\rho}, \tag{4.8}$$

$$P(M) = \frac{\alpha\theta}{n-1+\alpha\theta+\beta\rho}, \tag{4.9}$$

$$P(R) = \frac{\beta\rho}{n-1+\alpha\theta+\beta\rho}. \tag{4.10}$$

Il y a deux types de coalescence qui peuvent se produire dans nos généalogies. La première est une coalescence entre deux séquences absolument identiques de type i, C_i. La séquence résultante sera ainsi la même que ses deux descendantes (donc de type i aussi). Si on porte bien attention à l'équation (4.2), on remarque que c'est les probabilités d'observer un état τ connaissant l'état le précédant directement dans le passé $\tau+1$ dont on a besoin pour calculer $P(G)$. Pour observer une certaine génération H_τ, alors qu'on sait qu'elle diffère de la précédente $H_{\tau+1}$ par une coalescence C_i (dans la pratique, cela voudrait dire qu'un parent aurait transmis la même séquence génétique à deux de ses enfants), il faut que le premier évènement à avoir eu lieu soit une coalescence et également que ce soit une séquence de type i qui se soit multipliée. S'il y avait n_i séquences de type i à l'état τ, cela veut ainsi dire qu'il y en avait une de moins à l'état $\tau+1$ (et il y avait $n-1$ séquences au total). Ainsi, la probabilité conditionnelle recherchée est décrite par l'équation (4.11) :

$$P(H_\tau|H_{\tau+1} = H_\tau + C_i) = P(C)\frac{n_i-1}{n-1}. \tag{4.11}$$

Le deuxième type de coalescence qui peut exister est celle qui a lieu entre une séquence de type i et une de type j pour donner une séquence de type k, C_{ij}^k. Pour qu'un tel évènement soit possible, il faut que les marqueurs connus (les ancestraux) soient les mêmes et que la seule différence entre les deux séquences se trouve donc au niveau des marqueurs non ancestraux. La séquence k résultante

sera composée du matériel génétique commun à i et j en plus des marqueurs qui sont informatifs dans une séquence et pas dans l'autre. Si on sait que l'évènement séparant τ et $\tau + 1$ est une coalescence, la probabilité que ce soit i et j qui aient coalescé pour donner k s'exprime, lorsqu'on regarde du passé vers le présent, comme la probabilité qu'une séquence de type k ait été sélectionnée parmi les $n - 1$ séquences qu'il y avait dans $H_{\tau+1}$. Le nombre de séquences k présentes à la génération $\tau + 1$ est une de plus que le nombre de ces mêmes séquences que l'on observe à l'état τ sauf si une des séquences i ou j était déjà de type k (ce qui laisserait n_k inchangé dans le passé). En utilisant la notation du symbole de kronecker pour déterminer si i ou j sont de type k, on arrive à (4.12) :

$$P(H_\tau | H_{\tau+1} = H_\tau + C_{ij}^k) \ = P(C)\frac{n_k+1-\delta_{ik}-\delta_{jk}}{n-1}, \tag{4.12}$$

avec $\delta_{ik} = 1$ si $i = k$ et $\delta_{ik} = 0$ sinon ainsi que $\delta_{jk} = 1$ si $j = k$ et $\delta_{jk} = 0$ sinon.

Le troisième évènement qui peut être le lien entre H_τ et $H_{\tau+1}$ est un évènement de perte d'une mutation au marqueur m de la séquence i pour donner la séquence j, $M_i^j(m)$. Étant donné qu'avant la mutation, il y avait une séquence de type j de plus, mais que le nombre de séquences totales demeure inchangé (un évènement de mutation n'a pas d'effet sur la taille de l'échantillon), la probabilité que ce soit sur une séquence de type j qu'ait eu lieu la mutation est le rapport entre les deux. Sur une séquence, il y a en moyenne αp marqueurs où une mutation aurait pu avoir lieu. Ainsi la probabilité d'observer notre ensemble de séquences à la génération τ sachant que $M_i^j(m)$ est l'évènement transitoire entre elle et la génération qui la précède est (4.13) :

$$P(H_\tau | H_{\tau+1} = H_\tau + M_i^j(m)) \ = P(M)\frac{1}{\alpha p}\frac{n_j+1}{n}. \tag{4.13}$$

Le dernier type d'évènement qui peut arriver dans un graphe simulé est une recombinaison. Du passé vers le présent, cela veut dire qu'on aurait deux séquences

distinctes j et k qui s'interchangent une partie de leur matériel, ce qui fait en sorte que la séquence recombinée i contient la partie à gauche d'un point p d'un parent et la partie à droite de ce même point de l'autre parent. Nous considérons que ce point p est contenu dans un intervalle s de longueur r_s (rien n'oblige les distances entre nos marqueurs à être toutes égales). Du présent vers le passé, on modélise cet évènement par la division de i en j et k dont chacun va obtenir soit la partie gauche ou droite de i de l'intervalle s (le reste de sa séquence sera comblée par du matériel que l'on considère non ancestral), $R_i^{jk}(s)$. Considérons qu'on sait qu'un évènement de recombinaison a eu lieu. Comme il y a respectivement n_j et n_k séquences de type j et k à la génération τ (et donc $n_j + 1$ et $n_k + 1$ à la génération $\tau + 1$), il y a $(n_j + 1)(n_k + 1)$ couples de type j et k parmi les $(n + 1)n$ couples de séquences possibles (un évènement de recombinaison augmente de un la cardinalité de notre ensemble). De plus, sur le βr moyen disponible par séquence pour une recombinaison, la probabilité qu'elle ait eu lieu dans l'intervalle s est $r_s/\beta r$. Donc, la probabilité conditionnelle recherchée qui est associée à $R_i^{jk}(s)$ s'exprime comme dans (4.14) :

$$P(H_\tau | H_{\tau+1} = H_\tau + R_{ij}^k(s)) \quad = P(R)\frac{r_s}{\beta r}\frac{(n_j+1)(n_k+1)}{(n+1)n}. \tag{4.14}$$

La distribution, implémentée par Descary (2012), utilisée pour générer concrètement les graphes dans *DMap* et déterminer leur probabilité est celle développée par Fearnhead et Donnelly (2001). Afin de calculer la probabilité d'un graphe G, $Q(G)$, nous emploierons cette fois-ci un développement un peu différent de l'équation (4.2) :

$$\begin{aligned}
Q(G) &= Q(H_0, H_1, H_2, ..., H_{\tau^*}) \\
&= Q(H_{\tau^*} | H_0, H_1, ..., H_{\tau^*-1}) Q(H_0, H_1, ..., H_{\tau^*-1}) \\
&= Q(H_{\tau^*} | H_{\tau^*-1}) Q(H_0, H_1, ..., H_{\tau^*-1}) \\
&= Q(H_{\tau^*} | H_{\tau^*-1}) Q(H_{\tau^*-1} | H_{\tau^*-2}) Q(H_0, H_1, ..., H_{\tau^*-2})
\end{aligned}$$

$$
\begin{aligned}
&= \ldots \\
&= Q(H_0) \prod_{\tau=0}^{\tau^*-1} Q(H_{\tau+1}|H_\tau) \qquad (4.15) \\
&= \prod_{\tau=0}^{\tau^*-1} Q(H_{\tau+1}|H_\tau). \qquad (4.16)
\end{aligned}
$$

Étant donné que l'on ne génère que des graphes cohérents avec l'échantillon observé H_0, on peut affirmer $Q(H_0|G) = 1$, ce qui justifie le passage de (4.15) à (4.16). La preuve est faite dans Fearnhead et Donnelly (2001) que les probabilités conditionnelles de la distribution proposée optimale de Q, nécessaires au calcul de (4.16), sont :

$$Q(H_{\tau+1}|H_\tau) = P(H_\tau|H_{\tau+1}) \frac{\pi(H_{\tau+1})}{\pi(H_\tau)}. \qquad (4.17)$$

La distribution proposée Q est dite optimale, car les généalogies sont générées en ne laissant de côté aucune information connue sur H_0. $P(H_\tau|H_{\tau+1})$ correspondent aux probabilités conditionnelles démontrées plus haut pour chaque type d'évènement transitionnel alors que $\pi(H_\tau)$ se définit comme la probabilité qu'un échantillon aléatoire de séquences génétiques tiré dans une population soit le même que celui représenté dans H_τ, si on omet les marqueurs non ancestraux. $\pi(.)$ n'a pas de forme analytique connue, mais on peut la réécrire en définissant la distribution conditionnelle $\pi(.|H)$ comme étant la probabilité qu'une séquence d'un type connu ait été la dernière à s'ajouter à H dans le temps, connaissant les autres séquences contenues dans H. On est en mesure d'approximer celle-ci par une distribution $\hat{\pi}(.|H)$ (voir l'annexe A de Descary). Les détails de comment obtenir $\pi(H_{\tau+1})/\pi(H_\tau)$ à l'aide de $\pi(.|H)$ sont faits dans le chapitre 3 de Descary ; les trois quotients possibles sont exprimés dans les équations (4.18), (4.19) et (4.20) ci-bas :

$$\frac{\pi(H_{\tau+1} = H_\tau + C_{ij}^k)}{\pi(H_\tau)} = \frac{\pi(k|H_\tau - i - j)}{\pi(i|H_\tau - i)\pi(j|H_{\tau-i-j})}, \qquad (4.18)$$

$$\frac{\pi(H_{\tau+1} = H_\tau + M_i^j)}{\pi(H_\tau)} = \frac{\pi(j|H_\tau - i)}{\pi(i|H_\tau - i)}, \tag{4.19}$$

$$\frac{\pi(H_{\tau+1} = H_\tau + R_i^{jk})}{\pi(H_\tau)} = \frac{\pi(j|H_\tau - i)}{\pi(k|H_\tau - i + j)\pi(j|H_\tau - i)}. \tag{4.20}$$

Nous sommes donc maintenant en mesure de calculer $Q(G)$. Une fois un graphe généré, *DMap* utilise des vraisemblances et des tests du khi-carré et de permutation afin de faire une estimation de la position d'une mutation causant un certain phénotype. Étant donné que la façon de faire dans *DMapInteraction* est fortement inspirée de celle implémentée dans *DMap*, mais que celle de *DMap* ne s'applique pas directement dans un cadre plurigénique, nous n'introduirons dans ce mémoire que la première afin d'éviter des répétitions.

4.2 *DMap* Interaction

Étant donné le grand nombre de types d'interaction pouvant exister d'un point de vue biochimique, nous voulions avoir une méthode nous permettant d'obtenir une grande latitude au niveau de notre choix de modèles génétiques. C'est pour cela qu'en nous basant sur le modèle standard génétique à 3 pénétrances que l'on retrouve pour une maladie monogénique $F = (f_0, f_1, f_2)$, le modèle que nous avons choisi d'utiliser dans le cadre des interactions en est un à 9 pénétrances : $F = (f_{00}, f_{10}, f_{01}, f_{11}, f_{20}, f_{02}, f_{21}, f_{12}, f_{22})$. Chacun des f_{ij} correspond à la probabilité d'être affecté par la maladie conditionnellement au fait d'avoir i allèles mutants au premier SNP causal et j allèles mutants au deuxième SNP causal, $i, j \in \{0, 1, 2\}$.

La première étape majeure effectuée par *DMapInteraction* est identique en tout point à celle effectuée par *DMap*, notamment, on génère de nombreux graphes de recombinaison ancestraux cohérents avec nos données en faisant appel au processus de coalescence selon la distribution proposée par Fearnhead et Donnelly (2001) comme expliqué dans la section précédente.

L'étape suivante est de prendre chacun des graphes et de calculer différentes statistiques liées à celui-ci. Les couples de positions que peuvent prendre les mutations causant notre maladie sont discrets. En effet, pour chaque séquence génétique comportant L marqueurs de positions $\{x_1, x_2, ..., x_L\}$, les positions où chacune de ces mutations peut se situer se trouvent entre deux marqueurs successifs, mais nous les considérons pratiquement collées sur le marqueur les précédant, ce qui nous donne comme possibilités pour chaque marqueur les positions $\{y_1, y_2, ..., y_{L-1}\}$. En considérant tous les couples (y_i, y_j), $i, j \in \{1, ..., L-1\}$, cela nous donne un total de $(L-1)^2$ couples de positions possibles pour nos deux SNPs causaux. L'idée exploitée est que, étant donné la grande proximité entre x_i et y_i ($i \in \{1, ..., L-1\}$), pour chacune des positions y_i d'un SNP, nous pouvons lui associer le même arbre extrait de l'ARG que celui associé au marqueur x_i. De plus, étant donné que les mutations que l'on utilise sont non récurrentes, la mutation causale est apparue une fois dans l'histoire et n'a jamais remuté vers l'allèle du MRCA de notre échantillon de séquences. On modélise cela en affirmant que la mutation a dû apparaître sur une des branches de l'arbre du SNP et que toutes les séquences descendant de la première à hériter de la mutation dans l'histoire (la séquence du bas de cette branche) doivent aussi être porteuses de celle-ci. Étant donné que chaque paire de séquences correspond au matériel génétique d'une personne de notre échantillon, nous obtenons donc un génotype et un phénotype pour chaque individu. On teste donc toutes les possibilités de couples de branches pour toutes les paires (y_i, y_j) et calculons une vraisemblance et des tests du khi-deux pour chacune de celles-ci, comme nous le verrons dans la prochaine section.

Un exemple de ce processus est illustré dans la figure 4.2. On voit en premier lieu l'ARG, puis les deux arbres distincts qui peuvent être extraits de celui-ci : l'arbre du centre pour les deux premiers marqueurs, qui descendent du parent gauche lors de la recombinaison, et l'arbre du bas pour les deux derniers, qui descendent du

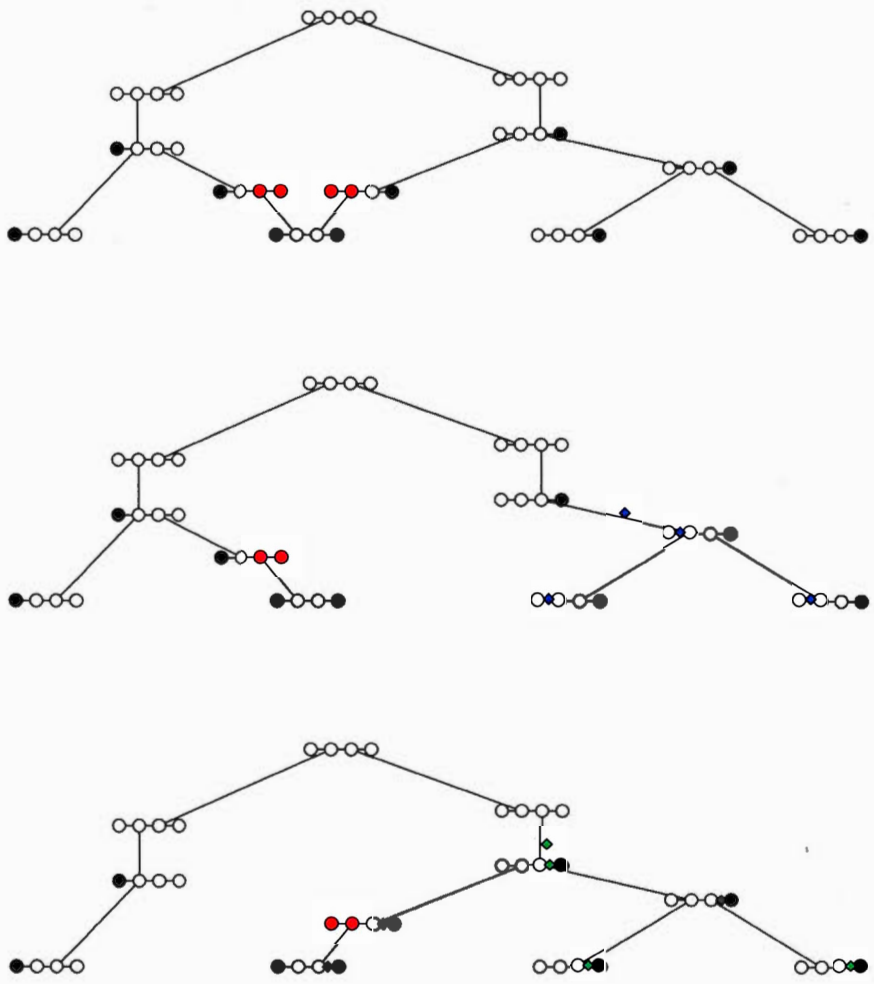

Figure 4.2: Exemple illustré de *DMapInteraction* une fois les graphes générés dans un cas où notre échantillon est de taille $n = 4$. L'image du haut est l'ARG complet, l'image du centre est l'arbre des 2 premiers marqueurs et l'image du bas celui des 2 derniers marqueurs.

parent droit. Ainsi nous avons quatre marqueurs dans notre exemple de positions $\{x_1, x_2, x_3, x_4\}$ et les neuf couples de positions de mutations testées seront donc

$(y_i, y_j), i, j \in \{1, 2, 3\}$ qui seront très près des trois premiers marqueurs. L'arbre utilisé pour les mutations de positions y_1 ou y_2 sera le premier et celui utilisé pour la mutation de position y_3 sera le second. Donc prenons l'exemple du couple (y_1, y_3).

Individu	Séquences	Phénotype	Mutations à la pos. 1	Mutations à la pos. 3
1	1 et 3	Non-malade	1	1
2	2 et 4	Malade	1	2

Tableau 4.2: Phénotypes et génotypes de l'échantillon de l'exemple de la figure 4.2.

Il faudrait bien sûr tester toutes les combinaisons de branches des deux arbres, mais nous illustrons seulement une d'entre elles. Nous voyons que selon les deux branches sélectionnées, les troisième et quatrième séquences héritent de la mutation à la position y_1, colorée en bleu, alors que toutes les séquences à l'exception de la première héritent de la mutation à la position y_3, colorée en vert. Considérant que la première et la troisième séquence appartiennent au premier individu (témoin) et que les deux autres appartiennent au deuxième individu (cas), cela nous donne les génotypes (1,1) et (1,2) respectivement comme on peut le voir dans le tableau 4.2. C'est cette information ajoutée à celle de toutes les autres combinaisons qui est utilisée pour les calculs de nos statistiques. En effet, étant donné que, dans cet exemple, chaque arbre a 10 branches, on aura $10^2 = 100$ combinaisons de branches à évaluer, et donc de tableaux de ce type à construire, pour chacune de nos neuf paires de SNPs causaux potentiels.

4.3 L'estimation

Afin d'estimer les positions où les mutations ont le plus de chances de se trouver, nous employons deux outils majeurs : des calculs de vraisemblance et des tableaux

du khi-carré pour chacun des couples de positions candidates.

4.3.1 La vraisemblance

Le premier outil que nous utilisons est la vraisemblance des différents couples de positions possibles pour nos mutations causant le phénotype observé. La vraisemblance est en fait la probabilité d'observer notre échantillon de n individus (et donc de $2n$ séquences génétiques), H_0, ainsi que le statut de chacun (malade ou pas) alors que nos deux mutations sont à des positions spécifiques, que nous notons c_{T_1} et c_{T_2}. L'échantillon de nos n phénotypes est indiqué par $\Phi = (\phi_1, ..., \phi_n)$. Les équations énoncées dans cette section sont adaptées à notre contexte à partir de Descary (2012) qui porte sur le cas d'une maladie monogénique. On a :

$$\begin{aligned} L(c_{T_1}, c_{T_2}) &= P(H_0, \Phi | c_{T_1}, c_{T_2}) \\ &= \int P(H_0, \Phi | G, c_{T_1}, c_{T_2}) P(G | c_{T_1}, c_{T_2}) dG \\ &= \int P(H_0 | \Phi, G, c_{T_1}, c_{T_2}) P(\Phi | G, c_{T_1}, c_{T_2}) P(G | c_{T_1}, c_{T_2}) dG \qquad (4.21) \\ &\approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(H_0 | \Phi, G^{(i)}, c_{T_1}, c_{T_2}) P(\Phi | G^{(i)}, c_{T_1}, c_{T_2}) P(G^{(i)} | c_{T_1}, c_{T_2})}{Q(G^{(i)})}. \end{aligned}$$

(4.22)

Étant donné l'infinité des généalogies possibles, le calcul de l'intégrale (4.21) est impossible à effectuer en pratique. Nous utilisons donc le principe d'échantillonnage pondéré afin d'approximer la vraisemblance exacte avec M graphes simulés comme exprimé dans l'équation (4.22). La densité alternative $Q(.)$ que nous avons employée dans *DMapInteraction* pour réaliser l'échantillonnage préférentiel est celle proposée par Fearnhead et Donnelly (2001) et a été introduite à l'équation (4.17) de la section 4.1. Les ARGs générés avec cette distribution étant tous construits à partir des données présentes dans H_0, on a $P(H_0 | \Phi, G^{(i)}, c_{T_1}, c_{T_2}) = 1$ pour tout $i \in \{1, ...M\}$. De plus, étant donné que

nous ne considérons pas que les deux mutations causales sont présentes dans nos séquences échantillonnées, le graphe généré ne dépend en aucun cas de leurs positions et, ainsi, $P(G^{(i)}|c_{T_1}, c_{T_2}) = P(G^{(i)})$. On obtient

$$L(c_{T_1}, c_{T_2}) \approx \frac{1}{M}\sum_{i=1}^{M}\frac{P(\Phi|G^{(i)}, c_{T_1}, c_{T_2})P(G^{(i)})}{Q(G^{(i)})} \tag{4.23}$$

$$= \frac{1}{M}\sum_{i=1}^{M}\left(P(\Phi|G^{(i)}, c_{T_1}, c_{T_2})\prod_{\tau=0}^{\tau^*-1}\left[\frac{P(H_\tau^{(i)}|H_{\tau+1}^{(i)})}{P(H_\tau^{(i)}|H_{\tau+1}^{(i)})\pi(H_{\tau+1}^{(i)})/\pi(H_\tau^{(i)})}\right]\right)$$

$$= \frac{1}{M}\sum_{i=1}^{M}\left(P(\Phi|G^{(i)}, c_{T_1}, c_{T_2})\prod_{\tau=0}^{\tau^*-1}\left[\frac{\pi(H_\tau^{(i)})}{\pi(H_{\tau+1}^{(i)})}\right]\right). \tag{4.24}$$

Afin de déterminer la probabilité d'avoir obtenu l'échantillon de cas et de témoins que nous observons, supposant un graphe potentiel résumant la généalogie de notre échantillon ainsi que deux positions de mutation, comme illustré dans la section précédente, nous considérons que les mutations ont dû apparaître à un certain moment dans l'histoire généalogique de notre échantillon et donc les déposons sur toutes les combinaisons de deux branches possibles des arbres liés aux mutations potentielles. $B_{Ac_{T_i}}$ est l'ensemble des branches de l'arbre associé à la position c_{T_i}, $(i \in \{1, 2\})$. Il résulte que

$$\begin{aligned}
P(\Phi|G, c_{T_1}, c_{T_2}) &= \sum_{b_1 \in B_{Ac_{T_1}}}\sum_{b_2 \in B_{Ac_{T_2}}} P(\Phi, b_1, b_2|G, c_{T_1}, c_{T_2}) \\
&= \sum_{b_1 \in B_{Ac_{T_1}}}\sum_{b_2 \in B_{Ac_{T_2}}} \frac{P(\Phi, b_1, b_2, G, c_{T_1}, c_{T_2})}{P(G, c_{T_1}, c_{T_2})} \\
&= \sum_{b_1 \in B_{Ac_{T_1}}}\sum_{b_2 \in B_{Ac_{T_2}}} \frac{P(\Phi, b_1, b_2, G, c_{T_1}, c_{T_2})}{P(b_1, b_2, G, c_{T_1}, c_{T_2})}\frac{P(b_1, b_2, G, c_{T_1}, c_{T_2})}{P(G, c_{T_1}, c_{T_2})} \\
&= \sum_{b_1 \in B_{Ac_{T_1}}}\sum_{b_2 \in B_{Ac_{T_2}}} P(\Phi|G, c_{T_1}, c_{T_2}, b_1, b_2)P(b_1, b_2|G, c_{T_1}, c_{T_2}).
\end{aligned} \tag{4.25}$$

Pour ce qui est du premier élément du produit dans chaque terme sommé au sein de l'équation (4.25), il représente la probabilité d'observer les phénotypes de notre

échantillon, considérant le graphe, deux positions de mutations fixées ainsi que deux branches où ces dites mutations ont pu apparaître. Nous considérons notre modèle à 9 pénétrances connu et l'employons dans ce calcul. En effet, tel que vu à la section précédente, en sachant où une mutation est apparue dans l'arbre généalogique, nous savons exactement qui en a hérité (et donc le nombre d'allèles mutants à chacune de nos deux positions pour chaque individu) et leur statut. Pour une personne ayant la configuration (k,l), sa probabilité d'avoir développé la maladie est $P(\phi_j = 1|G, c_{T_1}, c_{T_2}, b_1, b_2) = f_{kl}$ et celle de l'avoir évitée est $P(\phi_j = 0|G, c_{T_1}, c_{T_2}, b_1, b_2) = 1 - f_{kl}$. En effectuant le produit de ces probabilités pour tous nos individus, nous obtenons l'équation (4.26) où $n_{cas,kl}$ est le nombre de personnes dans notre échantillon qui sont de configuration (k,l) et qui sont atteints et $n_{tem,kl}$ sont le nombre de cette configuration qui sont sains. On a :

$$\begin{aligned} P(\Phi|G, c_{T_1}, c_{T_2}, b_1, b_2) &= \prod_{j=1}^{n} P(\phi_j|G, c_{T_1}, c_{T_2}, b_1, b_2) \\ &= \prod_{k=0}^{2}\prod_{l=0}^{2} f_{kl}^{n_{cas,kl}}(1 - f_{kl})^{n_{tem,kl}}. \end{aligned} \tag{4.26}$$

L'échantillonnage que nous utilisons généralement pour nos données est stratifié étant donné que la prévalence de la maladie peut être assez faible, ce qui nous demanderait une taille trop grande si on voulait un nombre de malades adéquat dans un échantillon aléatoire simple. Nous appliquons donc un facteur correcteur à notre vraisemblance, qui est théoriquement définie pour un échantillon de ce dernier type. Dénotons par t_m la fréquence de la maladie dans notre population. Dans un échantillon aléatoire simple, la fréquence de la maladie au sein de celui-ci devrait être approximativement la même. Donc, considérons que $t_m = n_{cas}/n'$ où n' serait la taille d'échantillon nécessaire pour avoir la même proportion de malades dans celui-ci que dans la population (étant donné que c'est nos cas qui sont surreprésentés dans notre échantillon, c'est leur nombre qui est utilisé pour en estimer un plus grand). On peut déterminer, en partant du fait que $1 - t_m =$

$n_{tem'}/n'$ que

$$\begin{aligned} n_{tem'} &= (1-t_m)n' \\ &= (1-t_m)\frac{n_{cas}}{t_m}. \end{aligned} \tag{4.27}$$

Afin d'estimer chacun des $n_{tem',kl}$, on estime la proportion de nos témoins ayant chacune des configurations (k,l) à l'aide de nos observations, puis calculons ce qu'elle aurait représenté en terme d'individus si on avait eu $n_{tem'}$ témoins. En réutilisant la formule (4.27), on obtient

$$\begin{aligned} n_{tem',kl} &= \frac{n_{tem,kl}}{n_{tem}} n_{tem'} \\ &= n_{tem,kl}\frac{n_{cas}(1-t_m)}{n_{tem}t_m}, k,l \in \{0,1,2\}. \end{aligned} \tag{4.28}$$

On remplace ensuite les $n_{tem,kl}$ de l'équation (4.26) par les $n_{tem',kl}$ de l'équation (4.28), ce qui nous donne comme premier élément de la vraisemblance (4.25) :

$$P(\Phi|G, c_{T_1}, c_{T_2}, b_1, b_2) = \prod_{k=0}^{2}\prod_{l=0}^{2} f_{kl}^{n_{cas,kl}}(1-f_{kl})^{n_{tem',kl}}. \tag{4.29}$$

En ce qui a trait au second et dernier élément du produit de l'équation (4.25), nous cherchons à déterminer la probabilité que ce soit sur deux branches b_1 et b_2 particulières qu'aient eu lieu les mutations. Comme nous considérons que la probabilité qu'une mutation ait eu lieu entre deux générations est proportionnelle au temps écoulé entre ces deux générations uniquement (d'où l'équation (4.30)), les branches sont indépendantes les unes des autres (le processus de coalescence étant un processus stochastique à accroissements indépendants). De cette propriété découle l'équation (4.31) :

$$\begin{aligned} P(b_1, b_2|G, c_{T_1}, c_{T_2}) &= P(b_1|G, c_{T_1}, c_{T_2})P(b_2|G, c_{T_1}, c_{T_2}) \\ &= P(b_1|G, c_{T_2})P(b_2|G, c_{T_2}), \end{aligned} \tag{4.30}$$

$$P(b_i|G, c_{T_i}) = \frac{|b_i|}{\sum_{w \in B_{Ac_{T_i}}} |w|}, i \in \{1,2\}. \tag{4.31}$$

4.3.2 Le khi-carré

Le second outil utilisé est une statistique du khi-deux calculée pour les différents couples de positions possibles pour nos mutations recherchées. On peut se poser la question : pourquoi utiliser le khi-carré alors qu'on a déjà la vraisemblance ? La justification est que cette technique est plus applicable dans un cas réel. Effectivement, dans la recherche médicale, on ne connaît pas avec exactitude le modèle de pénétrance génétique, ce qui rend le choix de la statistique du khi-deux assez intéressante comme cette dernière ne demande pas nécessairement une connaissance préalable de ce modèle.

Notre première intuition pour calculer $\chi^2(b_1, b_2)$ a été de le faire à partir d'un tableau 9x2 pour chaque combinaison de deux branches des arbres respectifs des deux positions de mutation testées, b_1 et b_2. Cette statistique a donc $(9-1)(2-1)$ =8 degrés de liberté. La répartition naturelle de ce tableau est de faire correspondre aux deux colonnes les statuts malade et non-malade alors que les lignes représentent pour leur part les 9 configurations (k, l) possibles des individus. Nous choisissions ensuite, pour chaque combinaison de c_{T_1} et c_{T_2}, la combinaison de deux branches nous donnant la valeur maximale pour le khi-carré et stockons en mémoire cette valeur. Le choix du maximum a été inspiré par ce qui a été fait par Descary (2012) dans *DMap*, elle-même guidée par Minichiello et Durbin (2006) dans leur méthode de cartographie génétique *Margarita*. Une moyenne est ensuite faite pour les M graphes utilisés comme on peut le constater dans l'équation (4.32). Le couple (c_{T_1}, c_{T_2}) ayant la valeur maximale de la statistique (et donc, par le même fait, la p-valeur minimale) correspond à notre estimation.

$$\chi^2(c_{T_1}, c_{T_2}) = \frac{1}{M} \sum_{i=1}^{M} \max_{b_1 \in B_{Ac_{T_1},i}, b_2 \in B_{Ac_{T_2},i}} \chi^2(b_1, b_2). \tag{4.32}$$

4.4 Les problèmes et solutions informatiques de *DMapInteraction*

Lors de l'implémentation en C++ de la méthodologie de *DMapInteraction*, nous nous sommes retrouvés confrontés à plusieurs défis computationnels auxquels nous avons cherché des solutions.

Le défi majeur fut au niveau de la vitesse d'exécution du programme. Étant donné la grande taille de nos ARGs et l'ordre quadratique des couples de branches à évaluer dans la partie inférence de *DMapInteraction*, nous avons cherché un moyen d'avoir une bonne approximation de notre vraisemblance et de notre khi-carré tout en réduisant nos temps de calcul, qui s'avéraient démesurés lors de nos premiers essais. Nous avons ainsi choisi d'utiliser une information obtenue facilement à partir de notre échantillon, soit la fréquence de la mutation au sein de celui-ci lorsque la mutation se produit sur une branche donnée. Connaissant μ, le paramètre représentant la fréquence de chacune des mutations dans la population (nous les considérons égales), nos pénétrances f_{ij} $(i,j \in \{0,1,2\})$ et t_m, la fréquence de notre maladie dans la population, nous sommes en mesure d'estimer la fréquence de celles-ci dans l'échantillon, que nous notons μ_{e,T_1} et μ_{e,T_2} par les calculs ci-dessous. Mentionnons que T_1 et T_2 sont les variables aléatoires exprimant le nombre d'allèles mutants pour un individu aux positions candidates et que nous supposons que nous sommes dans un cas où T_1 et T_2 sont indépendants.

$$\begin{aligned} F_{i,T_1} &= P(\phi = 1 | T_1 = i) \\ &= P(\phi = 1|(i,0))P(T_2 = 0) + P(\phi = 1|(i,1))P(T_2 = 1) + P(\phi = 1|(i,2))P(T_2 = 2) \\ &= f_{i0}(1-\mu)^2 + 2f_{i1}\mu(1-\mu) + f_{i2}\mu^2, i \in \{0,1,2\}, \end{aligned} \tag{4.33}$$

$$\begin{aligned} F_{i,T_2} &= P(\phi = 1 | T_2 = i) \\ &= P(\phi = 1|(0,i))P(T_1 = 0) + P(\phi = 1|(1,i))P(T_1 = 1) + P(\phi = 1|(2,i))P(T_1 = 2) \\ &= f_{0i}(1-\mu)^2 + 2f_{1i}\mu(1-\mu) + f_{2i}\mu^2, i \in \{0,1,2\}. \end{aligned} \tag{4.34}$$

$$\begin{aligned}
P(T_j = 1|\phi = 1) &= \frac{P(\phi = 1|T_j = 1)P(T_j = 1)}{P(\phi = 1)} \\
&= \frac{F_{1,T_j} \cdot 2\mu(1-\mu)}{t_m}, \\
P(T_j = 1|\phi = 0) &= \frac{(1 - F_{1,T_j}) \cdot 2\mu(1-\mu)}{1 - t_m}, \\
P(T_j = 2|\phi = 1) &= \frac{P(\phi = 1|T_j = 2)P(T_j = 2)}{P(\phi = 1)} \\
&= \frac{F_{2,T_j} \cdot \mu^2}{t_m}, \\
P(T_j = 2|\phi = 0) &= \frac{(1 - F_{2,T_j}) \cdot \mu^2}{1 - t_m}.
\end{aligned} \tag{4.35}$$

Afin de déterminer μ_{e,T_i} ($i \in \{1, 2\}$), nous estimons le nombre d'individus hétérozygotes (dont les deux allèles sont distincts) et le nombre d'individus homozygotes (dont les deux allèles sont identiques) porteurs pour une position, puis déterminons donc le nombre de séquences porteuses théoriques et divisons par le nombre de séquences totales afin d'avoir la proportion de séquences porteuses.

$$\mu_{e,T_j} = \frac{\sum_{i=1}^{2} P(T_j = i|\phi = 1) \cdot i \cdot n_{cas} + \sum_{i=1}^{2} P(T_j = i|\phi = 0) \cdot i \cdot n_{tem}}{2n}. \tag{4.36}$$

Ainsi, avec nos estimations des fréquences des mutations dans notre échantillon, nous établissons un critère de rejet pour les branches à utiliser dans les calculs de nos deux statistiques. En effet, si en déposant la mutation sur une branche de l'arbre de la position j, nous nous rendons compte que la proportion de gens qui en héritent est plus loin de μ_{e,T_j} qu'un paramètre ϵ au choix, nous ne calculons pas les vraisemblances et tableaux de khi-carré associés à tous les couples comprenant cette branche. Cela améliore la vitesse des calculs significativement.

Le deuxième défi principal auquel nous nous sommes retrouvés confrontés dans l'exécution de *DMapInteraction* est apparu dans l'utilisation du khi-carré. Les

conditions de ce test ne sont en général par respectées par les échantillons que nous utilisons. En effet, étant donné les tailles modestes de ceux-ci (qui ne peuvent être augmentées sans entraîner instantanément une lourdeur dans les calculs), il était possible d'avoir des cases vides dans nos tableaux de khi-carré. Cela arrive spécialement pour les configurations $(2,1)$, $(1,2)$ et $(2,2)$, étant donné la faible occurrence dans la population de nos mutations.

La première solution que nous avons choisie d'utiliser est de remplacer nos cases vides par des 0.5, afin de ne pas se retrouver avec une erreur dans la division du calcul de la statistique tout en ne biaisant pas trop celle-ci.

La seconde façon dont on a abordé le problème est en se disant qu'on éviterait ce désagrément en ayant des tableaux plus petits, dans notre cas 2x2. Or, ayant 9 lignes dans notre tableau initial, il y a $(2^9 - 2)/2$ =255 façons de les répartir en 2 lignes. Nous nous sommes donc basés sur les modèles génétiques utilisés pour générer notre maladie afin de choisir certaines répartitions stratégiques. Ces modèles sont basés sur des types d'interaction pouvant exister biologiquement. Si on prend un exemple où une maladie nécessite d'avoir les 2 gènes afin de se développer, une discrimination logique serait d'avoir les configurations $(0,0)$, $(1,0)$, $(0,1)$, $(2,0)$ et $(0,2)$ dans la première ligne (donc tous les individus n'ayant pas les deux mutations à la fois) et les configurations $(1,1)$, $(1,2)$, $(2,1)$ et $(2,2)$ dans la deuxième ligne. Il est certain que, dans une application à des données réelles dans un contexte de recherche clinique, nous ne serons pas en mesure de connaître le type d'interaction causant notre maladie, mais rien ne nous empêcherait de tester plusieurs types d'interaction et donc de tableaux dans le cadre d'une phase exploratoire, de garder les résultats pour chacun et de réduire par la suite la région sur laquelle la cartographie est faite pour ne contenir que les résultats préliminaires, ce qui nous permet d'avoir un échantillon plus grand pour cette seconde étape et donc, d'être en mesure d'utiliser notre tableau 9x2 à ce moment-là.

4.5 L'utilisation d'une généalogie connue

L'algorithme de *DMapInteraction* est divisé en deux sections majeures : la génération d'arbres de recombinaison ancestraux à partir d'un échantillon et l'estimation des deux positions des mutations causales à partir de ces graphes de recombinaison ancestraux. La distribution proposée par Fearnhead et Donnelly (2001) étant la norme dans la génération des ARGs depuis plusieurs années, on peut être en mesure de dire qu'elle a fait ses preuves. Nous voulions, par contre, tester la capacité de détection de nos vraisemblances et khi-carré. Nous avons donc décidé de faire des expériences en utilisant la généalogie réelle de notre échantillon.

Le logiciel permettant la génération de notre population, *fastsimcoal2* (Excoffier et al (2013)), que nous introduirons dans le prochain chapitre, nous permet, tout en générant notre population de séquences génétiques, d'obtenir un fichier de format nexus décrivant l'évolution de celle-ci à travers le temps. Le format nexus, utilisé beaucoup en biologie, contient la description des différents arbres composant notre ARG, chacun en notation Newick. La façon de lire un arbre de ce type se comprend mieux à l'aide d'un exemple. L'arbre (A :1,(B :1.5,C :2.75)) :1.5) est illustré dans la figure 4.3.

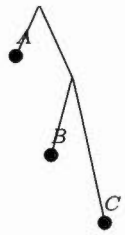

Figure 4.3: Illustration d'un arbre Newick.

Ainsi, les parenthèses les plus internes correspondent aux noeuds les plus bas dans la généalogie et plus un noeud est externe, plus il est haut. Les nombres associés

aux sommets sont les longueurs des branches (et donc les temps de coalescence dans notre situation). Dans notre exemple, cela signifierait que le temps entre la coalescence de B et C est de 1.5 (le minimum entre la longueur de la branche reliant B au noeud et celle reliant C au noeud). De même, le temps de coalescence entre A et le parent de B et C est 1.

La lecture d'un arbre Newick qui peut sembler simple de prime abord se complexifie rapidement lorsque la taille d'arbre grandit. Heureusement, le package « ape » en R contient la fonction « read.nexus » permettant de lire les fichiers de ce type et de stocker un arbre en un objet de classe phylo. La fonction « as.split » existant dans le package « phangorn » nous permet ensuite de générer pour chacun des arbres une matrice dont les colonnes sont les séquences de notre population et les lignes, toutes les branches formant l'arbre. Cette matrice est binaire et nous indique si chacun des individus descend d'une branche donnée ou pas (1 dans le cas où un individu descend de cette branche et 0 sinon). La matrice associée à l'arbre de la figure 4.2 serait donc celle ci-dessous :

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

La prochaine étape consiste à extraire la sous-matrice correspondant à notre échantillon, en gardant les colonnes les représentant ainsi qu'en éliminant les lignes ne contenant que des 0 après le retrait des colonnes des individus non échantillonnés. En effet, une ligne nulle indique qu'une branche n'a aucun descendant dans notre échantillon et ne fait donc pas partie de la généalogie de celui-ci. Par exemple, si on n'échantillonnait que B et C dans l'arbre de la figure 4.2, la sous-matrice associée à cet échantillon serait la suivante :

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Cette matrice nous donne toute l'information nécessaire pour donc déterminer les configurations génotypiques de chaque individu et l'associer à son phénotype. Les calculs de vraisemblance et de khi-carré se font ensuite comme dans la section précédente à l'exception près qu'on n'a maintenant qu'un seul graphe dans nos calculs et que celui-ci a un poids de 1 étant donné qu'on sait avec certitude qu'il représente l'histoire génétique véridique de notre échantillon.

CHAPITRE V

SIMULATIONS ET RÉSULTATS

Nous avons présenté dans le chapitre 4 *DMapInteraction*, une nouvelle méthodologie ayant comme objectif de localiser les deux mutations influençant un certain phénotype. Dans ce chapitre, nous présentons et analysons certains résultats que nous avons obtenu à l'aide de celle-ci. Nous commencerons par introduire le procédé que nous avons utilisé pour produire des ensembles de données, puis nous testerons notre méthodologie sur ces données en s'assurant de les explorer sous plusieurs angles.

5.1 La création des données

Étant donné que la méthode développée est encore à une étape préliminaire, nous usons de données simulées afin de la tester. L'accès à des données provenant du séquençage de réels génomes humains peut être dispendieux et un tel investissement n'est pas nécessaire à cette étape dans le processus. Ainsi, utiliser des données artificielles est un avantage majeur d'un point de vue financier. Mais le plus important est que, en simulant nos propres ensembles de données, on a la possibilité d'avoir la position réelle des SNPs causaux que nous désirons localiser, ce qui nous permet de valider l'efficacité de notre méthodologie.

Le logiciel que nous utilisons afin de simuler les données, *fastsimcoal2*, a été déve-

loppé par Excoffier et al (2013). Celui-ci est très complet et a maintes utilités : de la simulation de populations de séquences génétiques à l'inférence de certains paramètres démographiques sur des populations relativement complexes. Dans notre cas, on estime des populations de séquences génétiques composées de SNPs selon des paramètres contenus dans un fichier .par : le nombre d'haplotypes, le nombre de locis de chaque séquence, le taux de recombinaison, le taux de mutation ainsi que la fréquence minimum des allèles mutants désirés dans la population. On précise également à l'exécution que l'on veut que les mutations respectent le modèle des sites infinis abordé dans le chapitre 3 et si on veut connaître la généalogie exacte qui a donné lieu à la population ou pas.

Pour générer des populations d'haplotypes, *fastsimcoal2* utilise une version légèrement modifiée de SMC'. SMC' est une variante développée par Marjoram et Wall (2006) de l'algorithme SMC. Pour sa part, l'algorithme SMC, dont le nom est l'acronyme de *sequentially Markov coalescent*, est une méthode créée par McVean et Cardin (2005) permettant de générer des approximations des graphes de recombinaison ancestraux de façon beaucoup plus efficace computationnellement. L'idée générale est d'utiliser le principe qu'un graphe de recombinaison ancestral peut être décomposé, pour chacun de ses marqueurs, en un arbre de coalescence où des mutations peuvent avoir lieu. Ainsi, au lieu de produire directement un ARG, l'algorithme génère un arbre pour chacun des groupes de marqueurs entre lesquels il y a recombinaison en partant de la gauche vers la droite, en s'assurant que les arbres sont compatibles ensemble pour former un graphe de recombinaison ancestral.

La population générée est représentée sous la forme d'une suite de vecteurs où chacun représente un haplotype. Chaque vecteur, ayant le même nombre de marqueurs équidistants, correspond à une suite de 0, les marqueurs ancestraux, et de 1, les marqueurs ayant subi une mutation depuis l'époque de l'ancêtre commun

le plus récent. *Sample*, le logiciel développé par Fabrice Larribe (2011) en C++, prend ensuite la relève en lisant la population créée par *fastsimcoal2* et en fabriquant l'échantillon sur lequel il nous est possible de faire des analyses statistiques.

Sample commence par mélanger les haplotypes pour former des couples de façon aléatoire. Un couple d'haplotypes correspondra à un individu. Ainsi, si on a généré une population de taille N avec *fastsimcoal2*, *Sample* créera $N/2$ individus, d'où l'importance d'exiger un N pair. Ensuite, il choisit deux SNPs dont la fréquence dans la population est près d'une valeur qui lui est entrée en paramètre. Ces deux SNPs sont les TIMs, qui on le rappelle, sont les marqueurs qui influencent la présence du phénotype d'intérêt. Il nous est donc maintenant possible de connaître, pour chaque TIM, le génotype de chacun des individus. Est-il porteur de 0, 1 ou 2 allèles mutants à chacun des deux allèles causaux ?

En utilisant ensuite le modèle à 9 pénétrances évoqué dans le chapitre 4, il est possible de simuler un statut cas ou témoin à chacun des membres de la population. En effet, nous avions défini f_{ij} comme étant la probabilité d'être malade conditionnellement au fait d'être porteur de i mutations au premier TIM et j mutations au second ($i, j \in \{0, 1, 2\}$). Donc pour chaque individu, *Sample* observe son génotype (i, j), génère $x \sim U(0, 1)$ et lui assigne la maladie seulement si $x \leq f_{ij}$.

La dernière étape majeure est l'échantillonnage. Il peut être aléatoire simple ou stratifié, au choix de l'utilisateur. S'il est aléatoire simple, on spécifie une taille d'échantillon n, et n individus sont donc sélectionnés au hasard au travers de la population en entier. Par contre, la majeure partie du temps, il est stratifié, les études cas-témoins étant fréquentes en génétique humaines pour assurer la présence d'assez de personnes présentant le phénotype dans l'échantillon. Dans ce cas, on spécifie à *Sample* un nombre de cas (n_{cas}) et un nombre de témoins

(n_{tem}) désiré et il les pige de façon aléatoire dans chacune des deux strates de la population.

Nous testons dans les sections suivantes *DMapInteraction* sur 5 modèles génétiques distincts afin de mesurer sa capacité de détection dans différents cas. Le tableau 5.1 donne les pénétrances pour chacun des modèles utilisés, auquel on assigne une lettre de A à E. Nous utiliserons cette terminologie pour les désigner pour la suite des choses.

Modèle	f_{00}	f_{10}	f_{01}	f_{11}	f_{20}	f_{02}	f_{21}	f_{12}	f_{22}
A	0.01	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
B	0.01	0.02	0.02	0.99	0.1	0.1	0.99	0.99	0.99
C	0.01	0.1	0.1	0.2	0.4	0.4	0.5	0.5	0.99
D	0.01	0.7	0.02	0.71	0.85	0.1	0.95	0.75	0.99
E	0.01	0.1	0.1	0.4	0.4	0.4	0.6	0.6	0.7

Tableau 5.1: Les 5 modèles génétiques employés pour les tests de ce chapitre.

Essayons de visualiser plus concrètement quel type de maladie est représentée par chacun des modèles génétiques du tableau 5.1. Pour ce qui est de A, c'est un modèle dit à pénétrance complète, c'est-à-dire que dès qu'on a une mutation à un des deux TIMs, on est presque certain d'être malade. B est un modèle où avoir au moins une mutation à chaque TIM nous assure quasiment d'être malade, mais où les chances sont faibles si un individu n'est porteur d'aucun allèle mutant à au moins un des marqueurs causaux. Le modèle C est tel que si on est homozygote à un des marqueurs, on a plus de chance d'être malade que si on est hétérozygote mais avec le même nombre d'allèles mutants au total. Le nombre d'allèles mutants totaux fait également augmenter la prévalence de la maladie. Il y a donc un effet

de dominance dans un phénotype de ce type. Dans le modèle D, le premier TIM a beaucoup plus d'influence sur le fait de développer la maladie que le deuxième. C'est l'unique modèle du lot où les pénétrances ne sont pas symétriques. Des pénétrances sont symétriques si $f_{ij} = f_{ji}$ pour tout i et j. Le dernier modèle, le E, définit une maladie où ce n'est que le nombre d'allèles mutants portés par une personne qui détermine sa chance d'être malade, peu importe la provenance de ceux-ci. Plus ce nombre est grand, plus il est probable de la développer. Les fréquences des mutations dans la population sont dans tous les cas près de 15%.

5.2 Le cas des graphes de recombinaison ancestraux simulés

Dans cette section, nous présenterons les résultats de *DMapInteraction* lorsque l'on génère des arbres de recombinaison ancestraux selon la distribution proposée par Fearnhead et Donnelly (2001).

Commençons par expliquer, à l'aide de la figure 5.1, comment s'y prendre pour lire les graphiques produits par *DMapInteraction* qui seront présentés dans cette section ainsi que dans la suite du mémoire. Dans celle-ci, les deux axes sont de longueur identique, cette longueur étant celle des séquences de nos échantillons. La couleur se trouvant à une position donnée (c_{T_1}, c_{T_2}) indique la valeur du logarithme de la vraisemblance calculée pour un couple de marqueurs aux positions c_{T_1} et c_{T_2}. Les graphiques portant sur les χ^2 seront pratiquement identiques, à la distinction que c'est la mesure de l'inverse du logarithme du χ^2 calculé pour le couple de marqueurs qui se trouvera à la position (c_{T_1}, c_{T_2}). Des rappels sur la procédure utilisée pour calculer les vraisemblances et les χ^2 seront faits un peu plus loin dans la section. Quatre droites permettent de signaler au lecteur les positions des 2 marqueurs causant la maladie, comme on peut le voir à la figure 5.1. Les droites pointillées indiquent la position du premier TIM sur les 2 axes et les droites pleines indiquent la position du deuxième TIM sur les 2 axes. Ce sont ainsi les deux

points où une droite pleine croise une droite pointillée qui indiquent les positions des couples de marqueurs causaux. Les points blancs marquent pour leur part sur les graphiques toutes les positions des couples pour lesquels une vraisemblance maximale ou une p-valeur de χ^2 minimale (et donc une inverse de logarithme de p-valeur maximale) ont été calculés. Par exemple, dans la figure 5.1, les couples pour lesquels la vraisemblance était maximale sont les deux où un des marqueurs est à une position entre 0 et 0.005 et l'autre à une position très près de 0.030.

Pour ce qui est des figures où un test d'association du khi-deux basé sur un tableau de contingence 9x2 tel que présenté au chapitre 2 est fait, elles se lisent de la même façon qu'au chapitre 2, c'est-à-dire que le croisement de lignes pleines indiquent le couple représentant les positions des deux TIMs et le croisement de lignes pointillées indiquent celui dont l'inverse du logarithme de la p-valeur est maximale. Dans la suite de ce texte, afin d'éviter la confusion entre les khi-deux de *DMapInteraction* et les khi-deux introduits au chapitre 2, nous utiliserons l'appellation χ^2_{DMapInt} pour désigner les premiers.

Dans cette section, nous avons utilisé l'équation introduite au chapitre 4 pour calculer les vraisemblances lorsque M graphes sont générés :

$$L(c_{T_1}, c_{T_2}) \approx \frac{1}{M} \sum_{i=1}^{M} \left(P(\Phi|G^{(i)}, c_{T_1}, c_{T_2}) \prod_{\tau=0}^{\tau^*-1} \left[\frac{\Pi(H_\tau^{(i)})}{\Pi(H_{\tau+1}^{(i)})} \right] \right), \qquad (5.1)$$

où

$$P(\Phi|G, c_{T_1}, c_{T_2}) = \sum_{b_1 \in B_{Ac_{T_1}}} \sum_{b_2 \in B_{Ac_{T_2}}} P(\Phi|G, c_{T_1}, c_{T_2}, b_1, b_2) P(b_1, b_2|G, c_{T_1}, c_{T_2}). \qquad (5.2)$$

Par contre, nous avons choisi d'ignorer le terme $P(b_1, b_2|G, c_{T_1}, c_{T_2})$, car nous nous sommes rendus compte après de nombreuses simulations que l'omission de ce terme diminuait un peu le temps de calcul tout en donnant des résultats qui

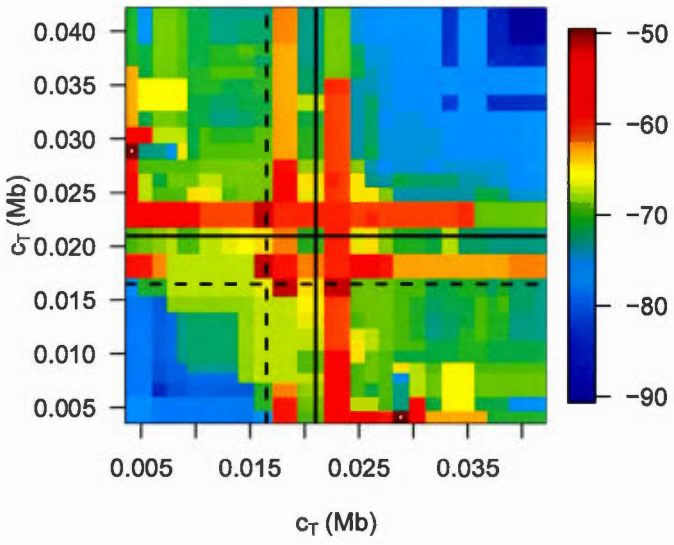

Figure 5.1: Exemple de figure produite par *DMapInteraction* illustrant la vraisemblance calculée pour chaque couple de points.

étaient meilleurs. Nous utilisons donc dans l'équation (5.1) :

$$P(\Phi|G, c_{T_1}, c_{T_2}) = \sum_{b_1 \in B_{Ac_{T_1}}} \sum_{b_2 \in B_{Ac_{T_2}}} P(\Phi|G, c_{T_1}, c_{T_2}, b_1, b_2), \tag{5.3}$$

où

$$P(\Phi|G, c_{T_1}, c_{T_2}, b_1, b_2) = \prod_{k=0}^{2} \prod_{l=0}^{2} f_{kl}^{n_{cas,kl}} (1 - f_{kl})^{n_{tem',kl}}. \tag{5.4}$$

Nous avons en effet observé, en utilisant les graphes représentant les généalogies véridiques de nos populations, que l'approximation de n_{tem} par le $n_{tem'}$ introduite dans le chapitre 4, qui rappelons le servait à donner une estimation du nombre de témoins qu'il y aurait dans un échantillon aléatoire simple si le nombre de cas de ce même échantillon était celui présent dans un échantillon stratifié, donnait de meilleurs résultats qu'en ne considérant que le nombre de cas et témoins réellement échantillonnés par échantillonnage stratifié. Des résultats comparant les deux approches seront présentés dans la section 5.3.

Pour ce qui est des statistiques du khi-deux, nous avons, comme mentionné au chapitre 4, calculé

$$\chi^2(c_{T_1}, c_{T_2}) = \frac{1}{M}\sum_{i=1}^{M} \max_{b_1 \in B_{Ac_{T_1},i}, b_2 \in B_{Ac_{T_2},i}} \chi^2(b_1, b_2) \tag{5.5}$$

en utilisant les tableaux d'association 9x2 où les colonnes coïncident avec le statut cas ou témoin et les lignes avec les différentes configurations génotypiques (i, j) que peut avoir un individu aux 2 TIMs.

Nous avons également utilisé la connaissance de nos modèles génétiques afin de créer des tableaux 2x2 appropriés pour l'étude des différents modèles génétiques comme nous l'avons proposé dans la section 4.4. Le tableau 5.2 présente la façon de regrouper les 9 couples de génotypes possibles en deux catégories pour étudier chacun des modèles génétiques A à E. Ainsi, dans la suite de ce chapitre, le χ^2_{DMapInt} de type i où $i \in \{1, 2, 3, 4, 5\}$ définira la catégorisation inspirée du modèle génétique j où $j \in \{A, B, C, D, E\}$ tels qu'ils sont associés dans le tableau 5.2. Nous sommes conscients qu'il pourrait y avoir d'autres façons de regrouper les lignes du tableau 9x2 en 2 catégories. Les cinq choix que nous avons faits nous paraissaient néanmoins intuitifs.

Type	Modèle	Première catégorie	Seconde catégorie
1	A	(1,0),(0,1),(1,1),(2,0),(0,2),(2,1),(1,2),(2,2)	(0,0)
2	B	(1,1),(2,1),(1,2),(2,2)	(0,0),(1,0),(0,1),(2,0),(0,2)
3	C	(2,0),(0,2),(2,1),(1,2),(2,2)	(0,0),(1,0),(0,1),(1,1)
4	D	(1,0),(1,1),(2,0),(2,1),(1,2),(2,2)	(0,0),(0,1),(0,2)
5	E	(1,1),(2,0),(0,2),(2,1),(1,2),(2,2)	(0,0),(1,0),(0,1)

Tableau 5.2: Les 5 types de χ^2 associés aux modèles génétiques.

Les figures 5.2 à 5.5 présentent les résultats de tous les tests mentionnés dans

cette section pour les modèles A à D. Le nombre de graphes simulés est M=750 et $n_{tem} = n_{cas}$=55. Les 7 premiers graphiques, tous produits par *DMapInteraction*, sont dans l'ordre (la lecture doit être faite de la gauche vers la droite et du haut vers le bas) : l'estimation par le maximum de vraisemblance, l'estimation par le χ^2_{DMapInt} déterminé à partir du tableau de contingence 9x2 et l'estimation par les χ^2_{DMapInt} de type 1 à 5 présentés dans le tableau 5.2. Le huitième et dernier graphique, celui situé en bas à droite de chacune des figures, estime le couple de marqueurs causaux par le test habituel d'association du khi-deux, qui se base sur un tableau de contingence 9x2. Notons que ce dernier n'est pas influencé par le nombre de graphes simulés M, car il ne calcule que l'association entre le nombre de mutations présentes aux 2 marqueurs formant un couple candidat et le fait d'être malade ou pas.

Notons que dans cette section, nous avons dû approximer sur de courtes séquences la diversité génomique présente sur des plus longues. En effet, plus les séquences génétiques d'un échantillon sont longues, plus la génération d'ARGs selon la distribution proposée par Fearnhead et Donnelly (2001) est coûteuse en temps, car les évènements de recombinaison sont plus nombreux. Comme mentionné aux chapitres 3 et 4, les évènements de recombinaison divisent une séquence en deux et augmentent ainsi la taille d'échantillon, ce qui nous éloigne de la convergence vers le MRCA. Par contre, des séquences trop courtes ont une pauvre diversité génétique à cause du déséquilibre de liaison fort présent entre deux marqueurs qui sont à proximité l'un de l'autre, comme vu au chapitre 2. Afin de pouvoir construire un nombre raisonnable d'ARGs, nous avons donc généré de relativement longues séquences avec *fastsimcoal2*, puis avons réduit arbitrairement les distances entre les marqueurs avant la lecture du fichier de données par *Sample*. La longueur de séquences considérée par notre *DMapInteraction* dans cette section est donc 0.04425 Mb.

Analysons maintenant les figures 5.2 à 5.5. On voit à la figure 5.2 que nous n'arrivons à détecter les TIMs avec aucune des 8 méthodes pour le modèle A. Il fallait s'attendre à cela pour les graphiques des χ^2_{DMapInt} de type 2 à 6 (donc les graphiques en position 4 à 7 si on numérote de gauche à droite et de haut en bas), car les catégorisations ne sont pas cohérentes avec le modèle génétique A. Mais les p-valeurs du χ^2_{DMapInt} général et de celui de type 1 pour les positions recherchées sont également plus grandes que pour des régions avoisinantes de tailles assez grandes. Pour ce qui est de la vraisemblance, on voit qu'il y a des couples très près des marqueurs causaux où la vraisemblance est d'un rouge assez foncé (donc grande sur l'échelle), par contre les maximums localisés sont en des points assez éloignés de ces couples. La différence entre les points localisés et les couples rouge foncé près de nos TIMs est très petite. On remarque que le χ^2, présent comme élément de comparaison, n'arrive pas non plus à détecter le couple de positions recherché.

À la figure 5.3, pour le modèle génétique B, nous n'arrivons pas non plus à détecter les TIMs d'aucune des huit manières, y compris avec la méthode que nous utilisons pour se comparer. Les couples de TIMs sont dans des régions orange foncé pour la vraisemblance, mais de vraisemblances plus grandes existent pour des couples comprenant le TIM 2 et d'autres TIMs dont les positions sont plus près de 0. Aucun des χ^2, que ce soit les χ^2_{DMapInt} ou le χ^2 usuel, n'a de faibles p-valeurs pour le couple cherché, en comparaison avec les autres couples testés.

Le maximum de vraisemblance appliqué à l'échantillon pour le test avec modèle génétique C a lieu pour deux points où une des cordonnées se situe entre le premier et le second TIM, mais encore une fois, nous n'arrivons pas à une estimation exacte. Pour ce qui est des χ^2_{DMapInt} et de la méthode comparative, aucun ne donne une localisation proche du couple à déceler. C'est ce qu'on observe dans la figure 5.4.

À la figure 5.5, on voit que pour le modèle génétique D, le χ^2_{DMapInt} à 9 lignes (deuxième image), le χ^2_{DMapInt} de type 1 (troisième image), le χ^2_{DMapInt} de type 3 (cinquième image) et le χ^2_{DMapInt} de type 5 (septième image) détectent tous des points dont un est près du premier TIM (on se souvient que le modèle D est celui où le TIM 1 est beaucoup plus influent au développement de la maladie que le TIM 2). Par contre, la détection de la dernière image est meilleure.

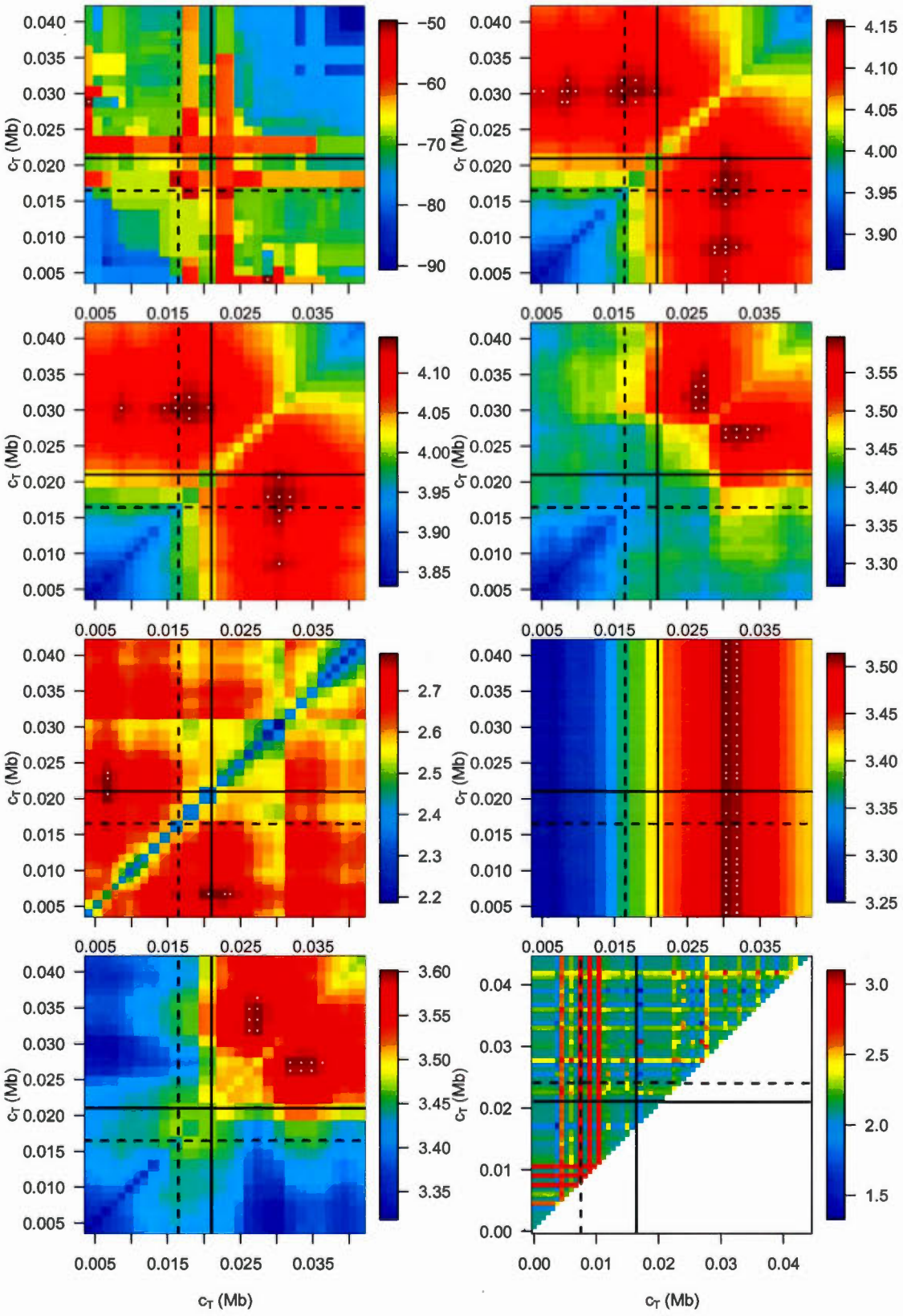

Figure 5.2: Modèle A. 55 cas et 55 témoins. Vraisemblance, les 6 χ^2_{DMapInt} et χ^2 standard.

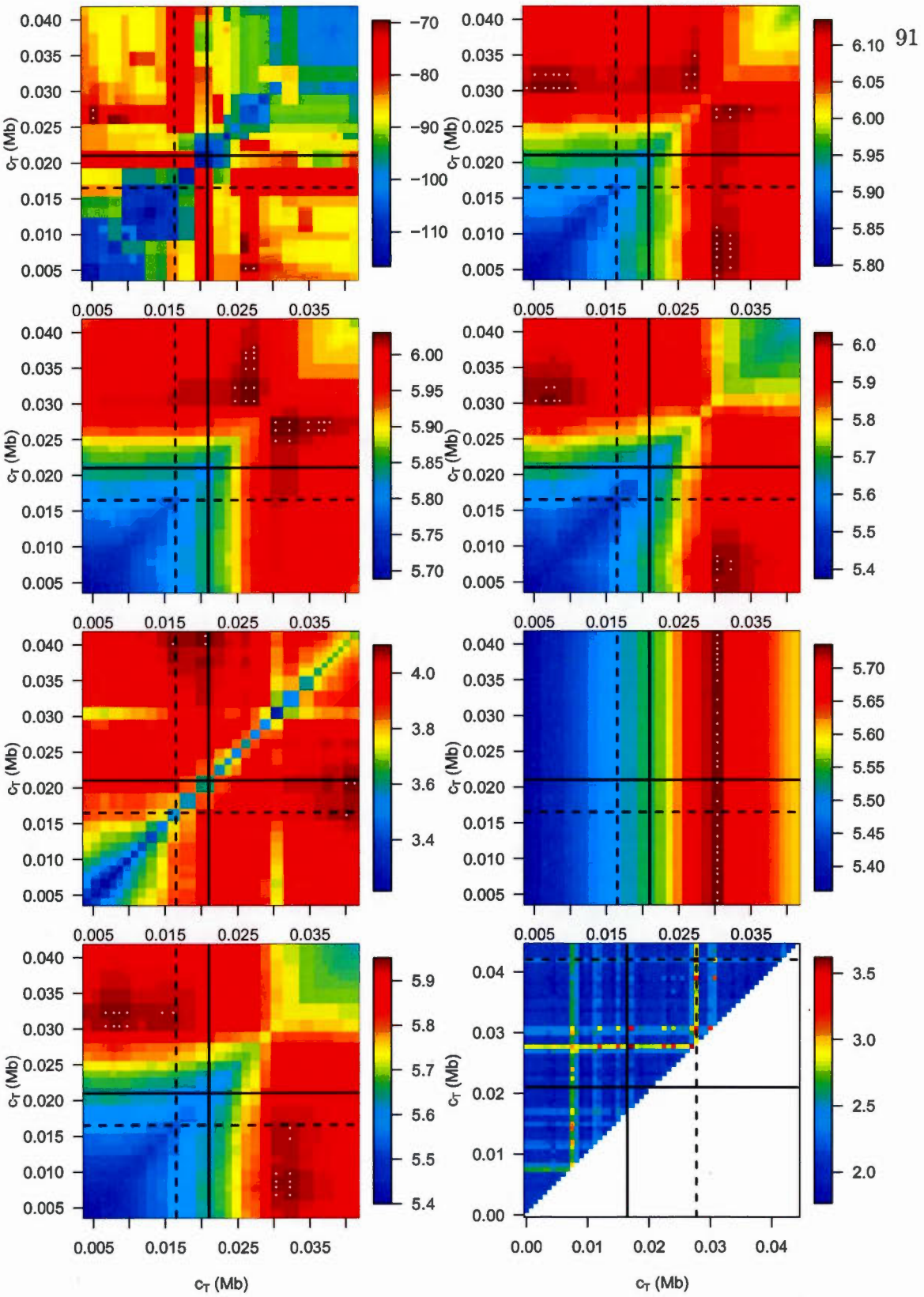

Figure 5.3: Modèle B. 55 cas et 55 témoins. Vraisemblance, les 6 χ^2_{DMapInt} et χ^2 standard.

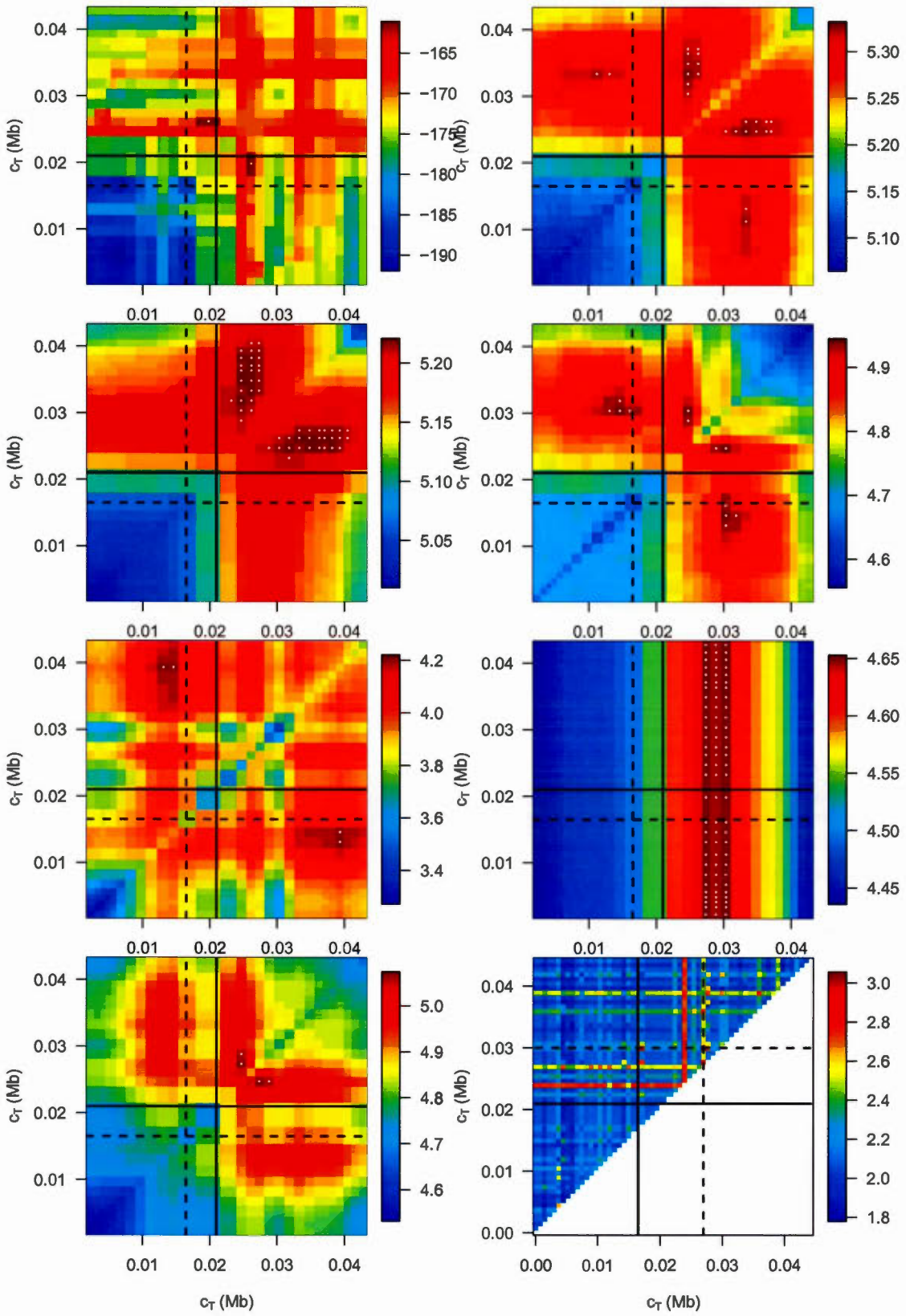

Figure 5.4: Modèle C. 55 cas et 55 témoins. Vraisemblance, les 6 χ^2_{DMapInt} et χ^2 standard.

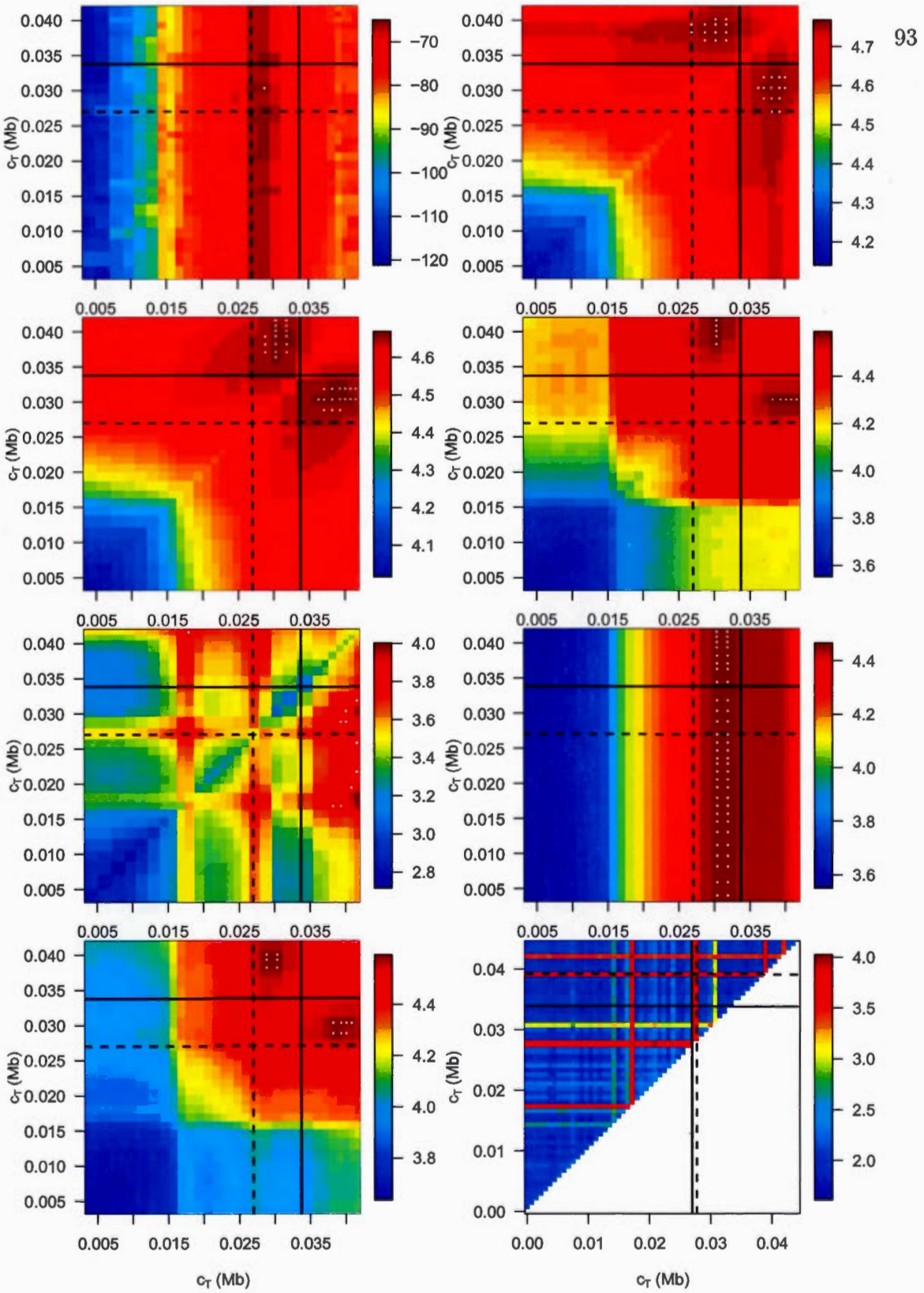

Figure 5.5: Modèle D. 55 cas et 55 témoins. Vraisemblance, les 6 χ^2_{DMapInt} et χ^2 standard.

Par les figures 5.6 et 5.7, nous voulons étudier, pour les modèles A et D, si l'augmentation de M, le nombre de graphes de recombinaison ancestraux simulés, améliore l'estimation des TIMs par le maximum de vraisemblance et avec la statistique χ^2_{DMapInt} principale, celle déterminée avec les tableaux 9x2, comme notre intuition nous le laisse croire. Nous avons donc fait varier M de 200 à 1700 pour le modèle A dans la figure 5.6 et de 200 à 2000 pour le modèle D dans la figure 5.7.

À la figure 5.6, on ne semble pas voir une amélioration de la détection lorsque le nombre d'ARGs simulés augmente quand nous faisons le test pour le modèle génétique A. En effet, les détections les plus près des positions des couples de TIMs par la vraisemblance ont lieu lorsque $M = 200$ et $M = 1000$. Elles sont par contre très éloignées lorsque $M = 750$ et $M = 1700$. On ne voit pas de différences flagrantes pour l'efficacité de l'estimation par le χ^2 pour aucune valeur de M.

Si on se base sur les résultats de la figure 5.7 pour le modèle D, on ne voit également aucun intérêt à augmenter le nombre de graphes simulés.

Étant donné que l'effet de l'augmentation de M ne semble être considérable, on pourrait être porté à croire que la méthode de Fearnhead et Donnelly est inefficace pour générer des ARGs qui estiment bien la réelle histoire de notre population. Par contre, cette méthode étant utilisée fréquemment dans le monde de la coalescence depuis plus d'une dizaine d'années, cela serait surprenant. Il faudrait ainsi être en mesure d'augmenter beaucoup plus le nombre d'ARGs simulés avant de conclure que ce nombre n'a aucun effet sur la précision de nos estimations, chose qui est impossible pour nous d'un point de vue informatique jusqu'à maintenant. On peut s'interroger aussi sur l'efficience de nos méthodes de détection si on considère que les ARGs générés approximent bien l'histoire généalogique de notre échantillon. C'est pour cela que nous laisserons de côté la simulation d'ARGs dans la prochaine section et nous concentrerons sur l'étude de nos techniques d'estimation.

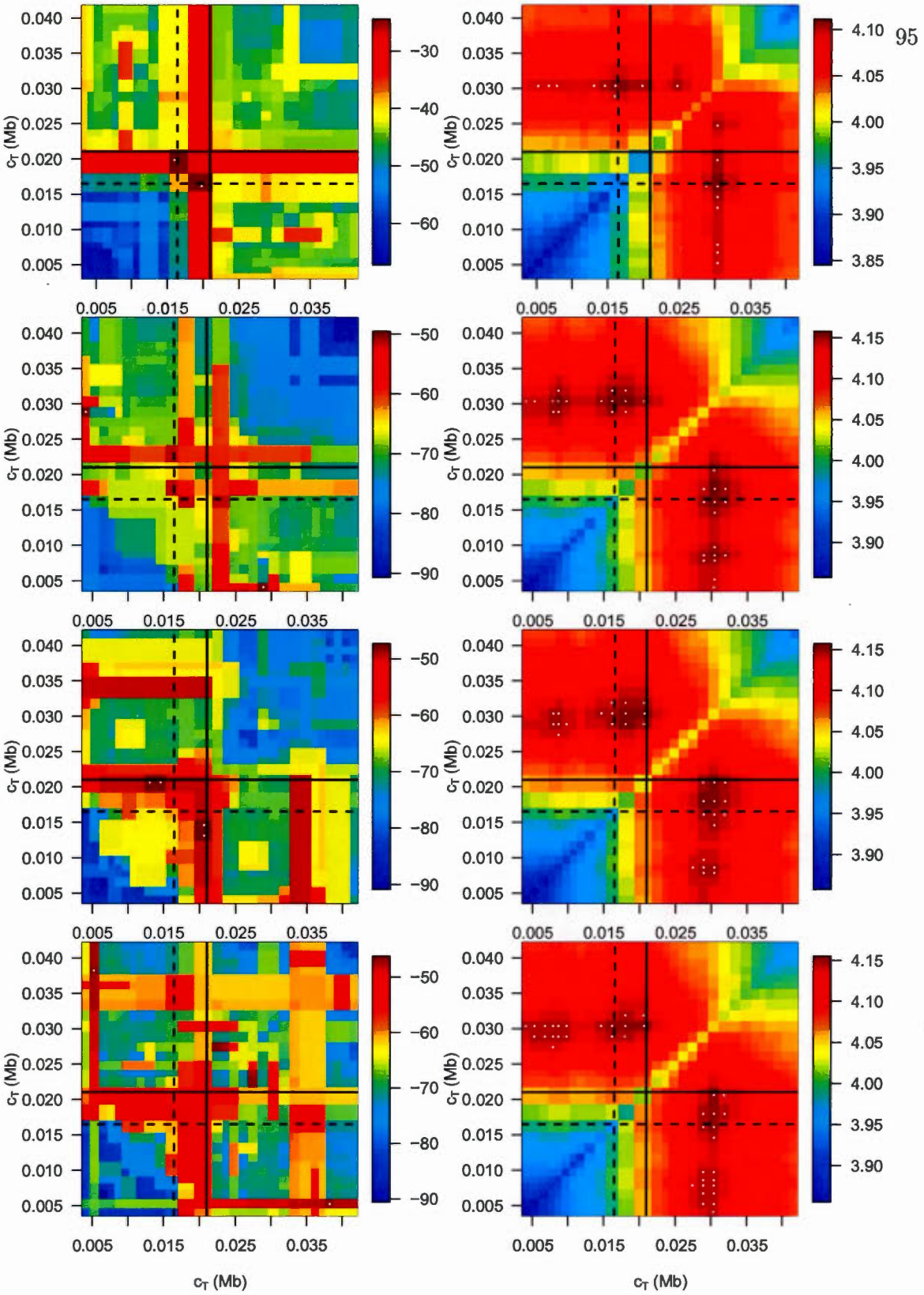

Figure 5.6: Modèle A. Vraisemblance à gauche et χ^2 à droite. Les nombres de graphes pour les lignes de haut en bas : 200, 750, 1000 et 1700.

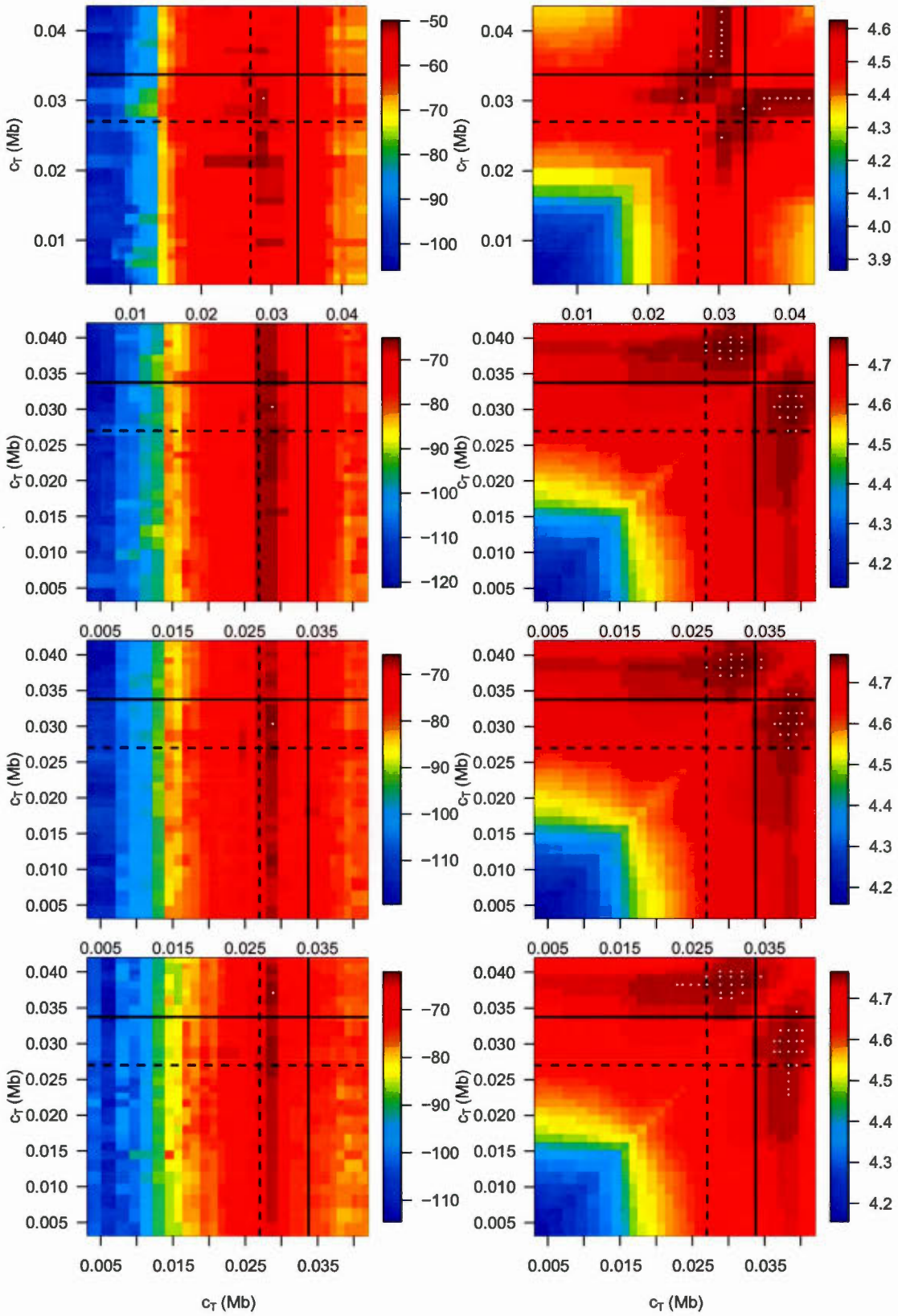

Figure 5.7: Modèle D. Vraisemblance à gauche et χ^2 à droite. Les nombres de graphes pour les lignes de haut en bas : 200, 750, 1000 et 1700.

5.3 Le cas où l'on connaît la généalogie

Comme nous l'avons vu dans la dernière section, nous avons de la difficulté à déterminer exactement la position de nos TIMs, que ce soit à l'aide de l'estimation par la vraisemblance ou celle par les différents χ^2. Comme mentionné dans la section 4.5, nous avons choisi d'évaluer notre méthodologie à partir de l'ARG représentant l'histoire réelle de la population étudiée. L'intérêt à procéder de cette façon, on le rappelle, est de mesurer de façon ciblée la capacité de détection de celle-ci sans considérer le risque d'erreur apporté par la génération de graphes par l'algorithme de Fearnhead et Donnelly. Dans la suite de ce chapitre, M est donc égal à 1 et la seule généalogie G sur laquelle on repose a une probabilité 1, ce qui fait que

$$L(c_{T_1}, c_{T_2}) \approx P(\Phi|G, c_{T_1}, c_{T_2}), \tag{5.6}$$

et

$$\chi^2(c_{T_1}, c_{T_2}) = \max_{b_1 \in B_{Ac_{T_1}}, b_2 \in B_{Ac_{T_2}}} \chi^2(b_1, b_2). \tag{5.7}$$

5.3.1 Le type d'échantillonnage

Le premier angle sous lequel nous voulons étudier notre méthode est celle du type d'échantillonnage. Ainsi, nous avons comparé les vraisemblances calculées sur un échantillon aléatoire simple de taille $n = 100$, un échantillon stratifié où $n_{cas} = 30$ et $n_{tem} = 30$, mais où on remplace n_{tem} par $n_{tem'}$ et un échantillon stratifié régulier où $n_{cas} = 30$ et $n_{tem} = 30$ respectivement. La figure 5.8 présente les résultats des trois estimations pour le modèle A à gauche et le modèle B à droite et la figure 5.9 les présente pour le modèle E.

Dans la figure 5.8, on observe que pour le modèle A comme pour le modèle B, la détection par le maximum de vraisemblance est parfaite lorsque nous utilisons un

échantillon aléatoire simple. Pour ce qui est du modèle A, on détecte également les TIMs lorsque nous utilisons l'échantillonnage stratifié, peu importe si le nombre de témoins utilisés dans l'évaluation de la vraisemblance est n_{tem} ou $n_{tem'}$. Pour le modèle B, on ne détecte pas le deuxième TIM, bien qu'on en soit très près, mais on voit que la détection est identique pour les deux cas où on utilise l'échantillonnage stratifié.

En ce qui a trait au modèle E, par contre, on ne détecte les TIMs que dans le cas où on utilise l'échantillonnage stratifié où on remplace n_{tem} par $n_{tem'}$, comme on peut le remarquer dans la figure 5.9. Avec l'échantillonnage stratifié régulier, on détecte le TIM 1 mais on est bien loin du TIM 2, alors qu'avec l'échantillonnage aléatoire, on est près du TIM 2 (sans être exactement dessus), mais assez loin du TIM 1.

Comme notre approximation de l'échantillonnage aléatoire dans un échantillonnage stratifié semble avoir un effet positif dans un cas, sans nuire à l'estimation dans les autres, c'est ce type d'échantillonnage que nous utiliserons pour la suite de ce texte.

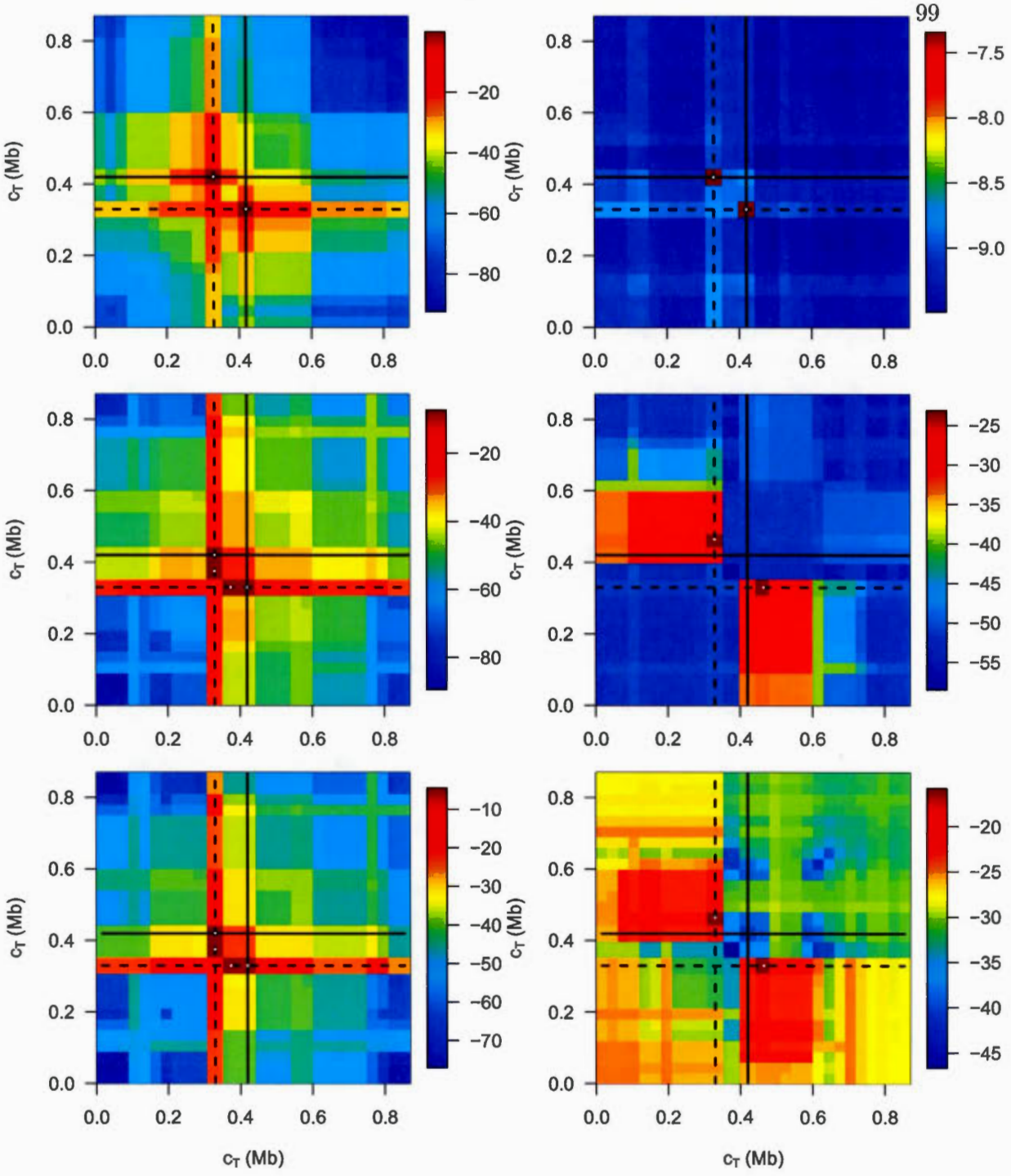

Figure 5.8: Modèle A (gauche) et modèle B (droite). Estimation par la vraisemblance. Échantillons aléatoires de taille n =100 (haut), échantillons stratifiés de 30 cas et 30 témoins où on utilise le $n_{tem'}$ (centre) et échantillons stratifiés réguliers de 30 cas et 30 témoins (bas).

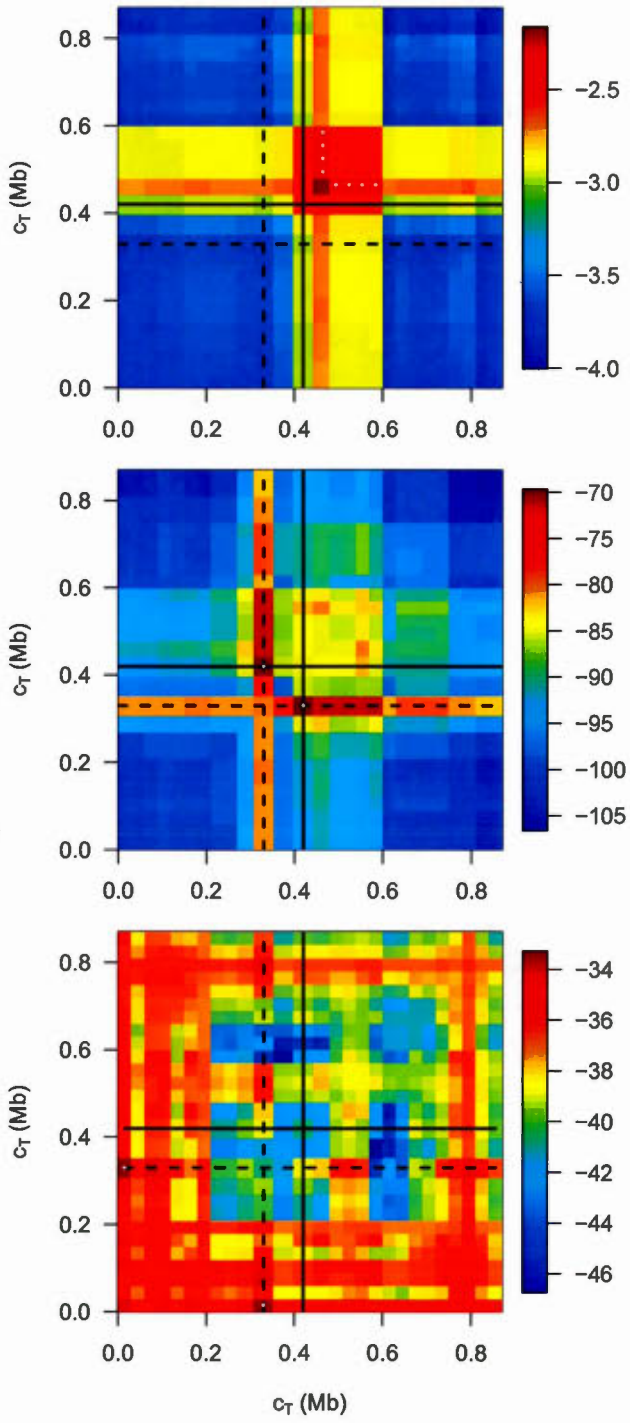

Figure 5.9: Modèle E. Estimation par la vraisemblance. Échantillons aléatoires de taille n =100 (haut), échantillons stratifiés de 30 cas et 30 témoins où on utilise le $n_{tem'}$ (centre) et échantillons stratifiés réguliers de 30 cas et 30 témoins (bas).

5.3.2 La taille d'échantillon

Le deuxième angle sous lequel nous voulons étudier notre méthode est celle des tailles d'échantillon. Nous voulions voir si augmenter les tailles des échantillons aurait un effet significatif sur la précision des estimations des TIMs. Nous avons donc comparé pour le modèle C les vraisemblances à la figure 5.10, les statistiques χ^2_{DMapInt} à 8 degrés de liberté de *DMapInteraction* à la figure 5.11 et les statistiques du χ^2 à 8 degrés de liberté régulières pour des tailles d'échantillons stratifiés variables. Dans chacun des cas, $n_{cas} = n_{tem}$, mais $n_{tem'}$ remplace n_{tem} dans les calculs. Les tailles choisies pour les tests varient de 30 cas et 30 témoins à 118 cas et 118 témoins pour les trois figures.

L'augmentation de la taille d'échantillon ne semble pas améliorer la détection du couple de TIMs lorsqu'on utilise la vraisemblance. En effet, dans la figure 5.10, on remarque que le point où la vraisemblance est maximale est le même pour toutes les 5 tailles d'échantillon utilisées. De plus, les points correspondants aux positions de nos TIMs se trouvent dans les 5 graphiques dans des zones où les couleurs sont assez froides, donc où les vraisemblances sont petites.

L'effet est différent à la figure 5.11, avec le χ^2_{DMapInt} basé sur le tableau 9x2. On détecte le TIM 1 pour les 5 tailles d'échantillon, mais on se rapproche du TIM 2 lorsque l'on a 75, 100 et 118 cas et témoins. De plus, lorsque l'on a 100 et 118 cas et témoins, nos estimations de couples de TIMs sont uniques.

À la figure 5.12, on observe que lorsqu'on utilise la méthode comparative, on ne détecte aucun des 2 TIMs pour aucune des tailles d'échantillon. On ne semble pas se rapprocher des marqueurs causaux non plus avec l'augmentation du nombre d'individus malades et en santé.

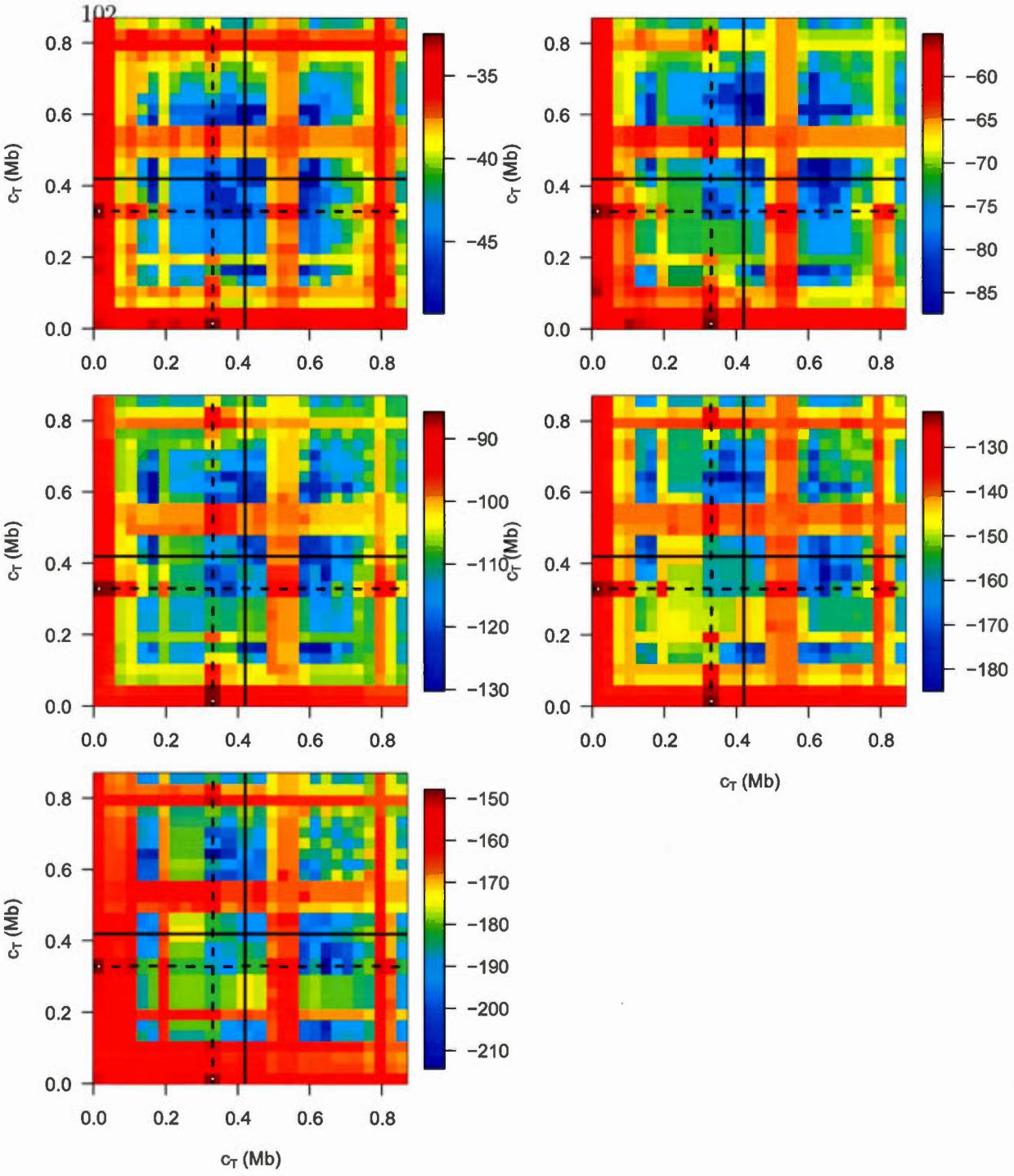

Figure 5.10: Modèle C. Estimation par la vraisemblance. Nombre de cas et témoins variant de gauche à droite et haut en bas : 30, 50, 75, 100, 118.

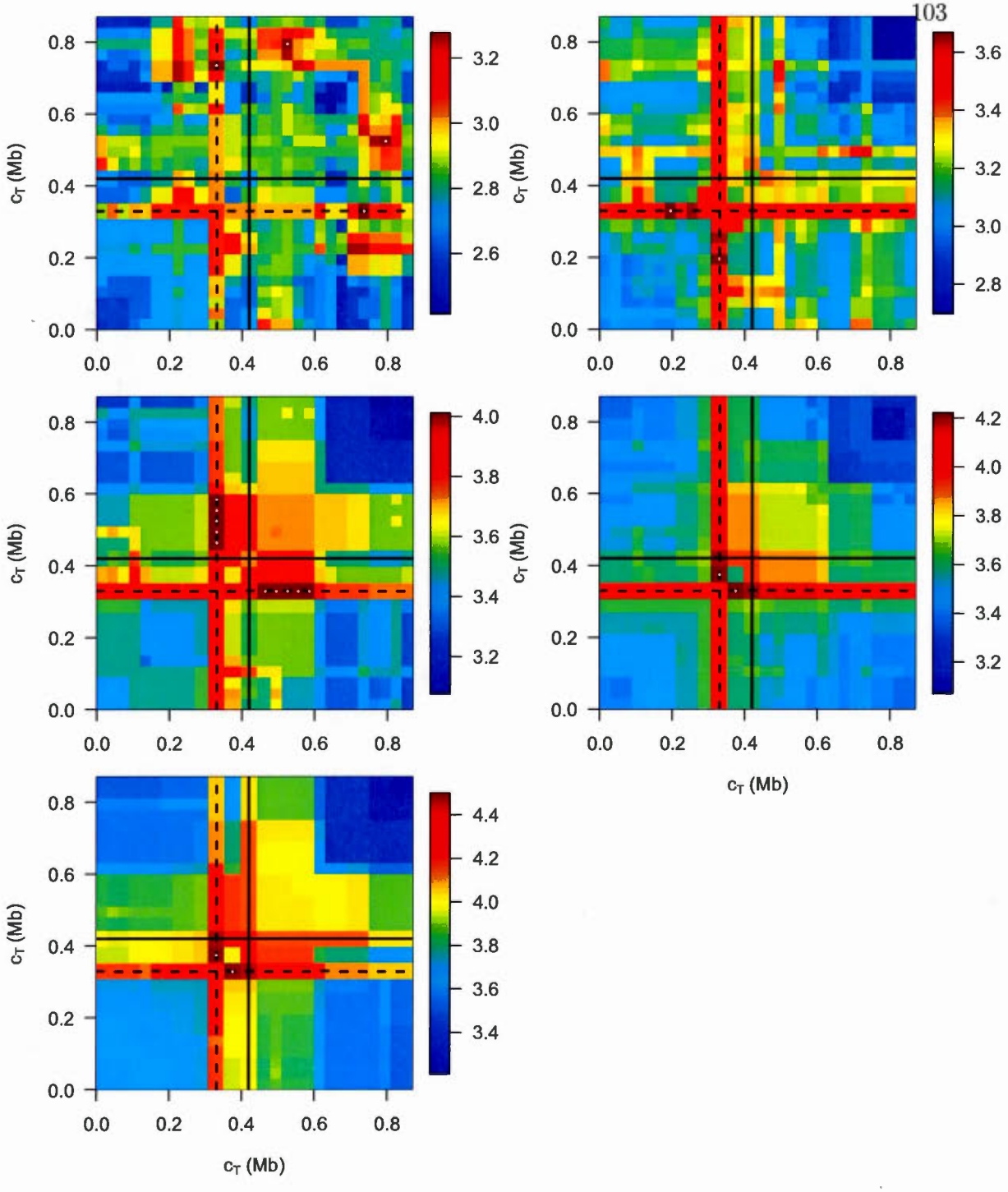

Figure 5.11: Modèle C. Estimation par le χ^2_{DMapInt}. Nombre de cas et témoins variant de gauche à droite et haut en bas : 30, 50, 75, 100, 118.

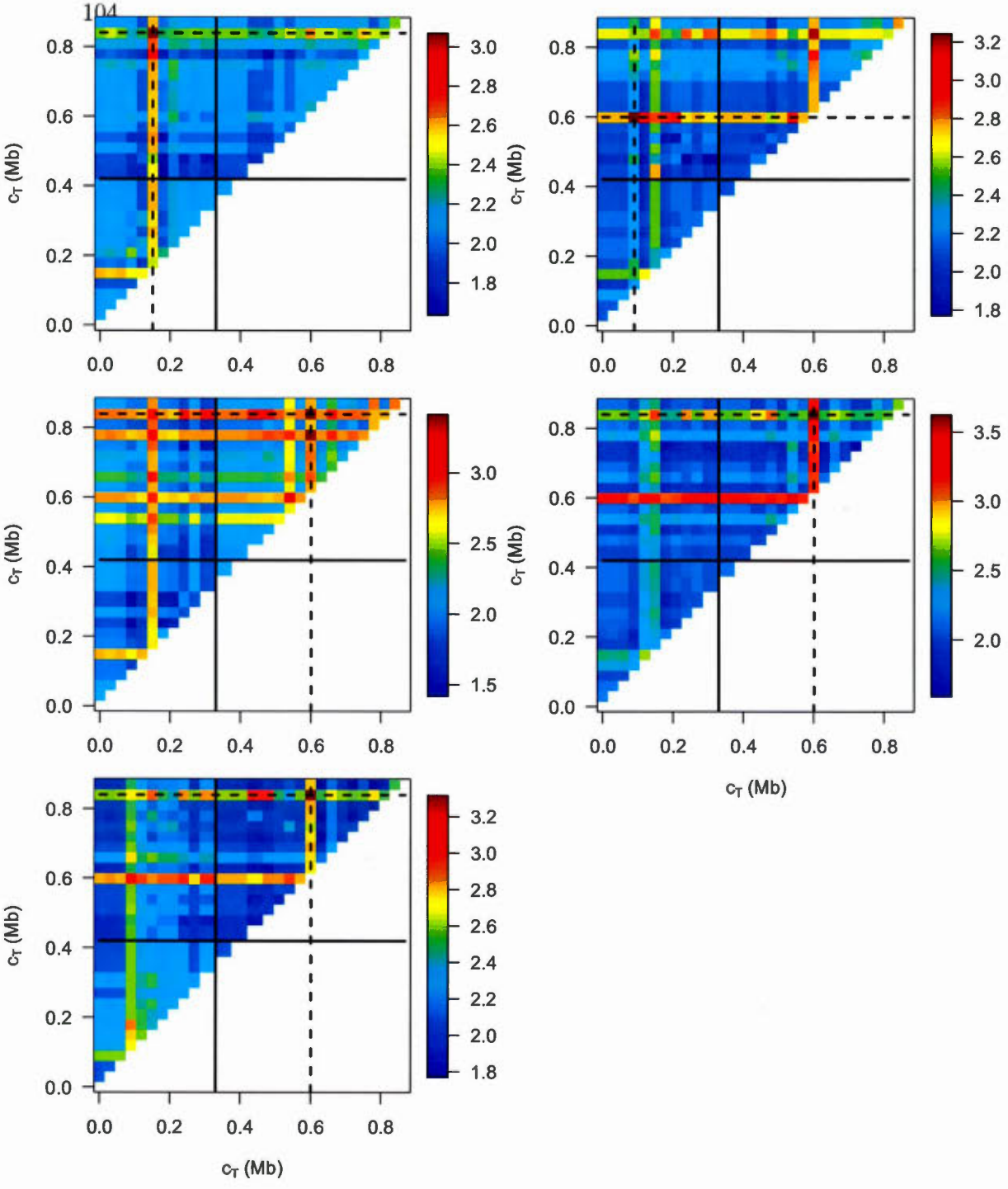

Figure 5.12: Modèle C. Estimation par le χ^2. Nombre de cas et témoins variant de gauche à droite et haut en bas : 30, 50, 75, 100, 118.

5.3.3 La connaissance du modèle génétique

Le dernier aspect présenté pour l'étude à l'aide de l'ARG réel de l'échantillon est l'utilisation, comme dans la section 5.2, de la connaissance du modèle génétique pour la création de tableaux de contingence 2x2 adaptés pour le calcul de χ^2_{DMapInt}. On rappelle que les 2 catégories pour le test du χ^2 adapté à chaque modèle ont été présentées dans le tableau 5.2. Les figures 5.13 à 5.15 illustrent donc pour chacun des modèles A à E, de haut en bas, l'estimation par la vraisemblance, l'estimation par le χ^2_{DMapInt} général et l'estimation par le χ^2_{DMapInt} de type 1 à 5 coïncidant avec le modèle étudié. Nous utilisons des échantillons où $n_{cas} = 80$ et $n_{tem} = 80$.

La figure 5.13 illustre que pour le modèle A, la vraisemblance, le χ^2 basé sur le tableau 9x2 ainsi que le χ^2 de type 1 détectent exactement le couple de TIMs recherchés. Pour ce qui est du modèle B, la vraisemblance et le χ^2 de type 2 le trouvent, mais le χ^2 ne tenant compte du modèle de pénétrances localise des points où le deuxième marqueur est à une position un peu plus grande que celle du TIM 2.

On voit à la figure 5.14 pour le modèle C que, comme pour le modèle B, la détection est correcte lorsque nous utilisons la vraisemblance ou le χ^2 tenant compte du modèle génétique, celui de type 3 dans ce cas-ci. Par contre, la détection est inexacte lorsque le χ^2 à 8 degrés de liberté est utilisé. Pour le modèle D, nous détectons le TIM 1 dans les trois cas. Par contre, nous n'arrivons pas à différencier les couples comprenant le TIM 2 de tous les autres couples dont fait partie le TIM 1 lorsque nous utilisons la vraisemblance et le χ^2 de type 4. Avec l'autre χ^2, nous détectons deux couples dont le TIM 1 fait partie avec un autre marqueur quelconque.

Finalement, à la figure 5.15, on arrive à détecter les TIMs à l'aide des trois techniques pour le modèle E.

On peut donc conclure que les χ^2 qui prennent en considération les pénétrances des diverses maladies s'avèrent être des outils intéressants.

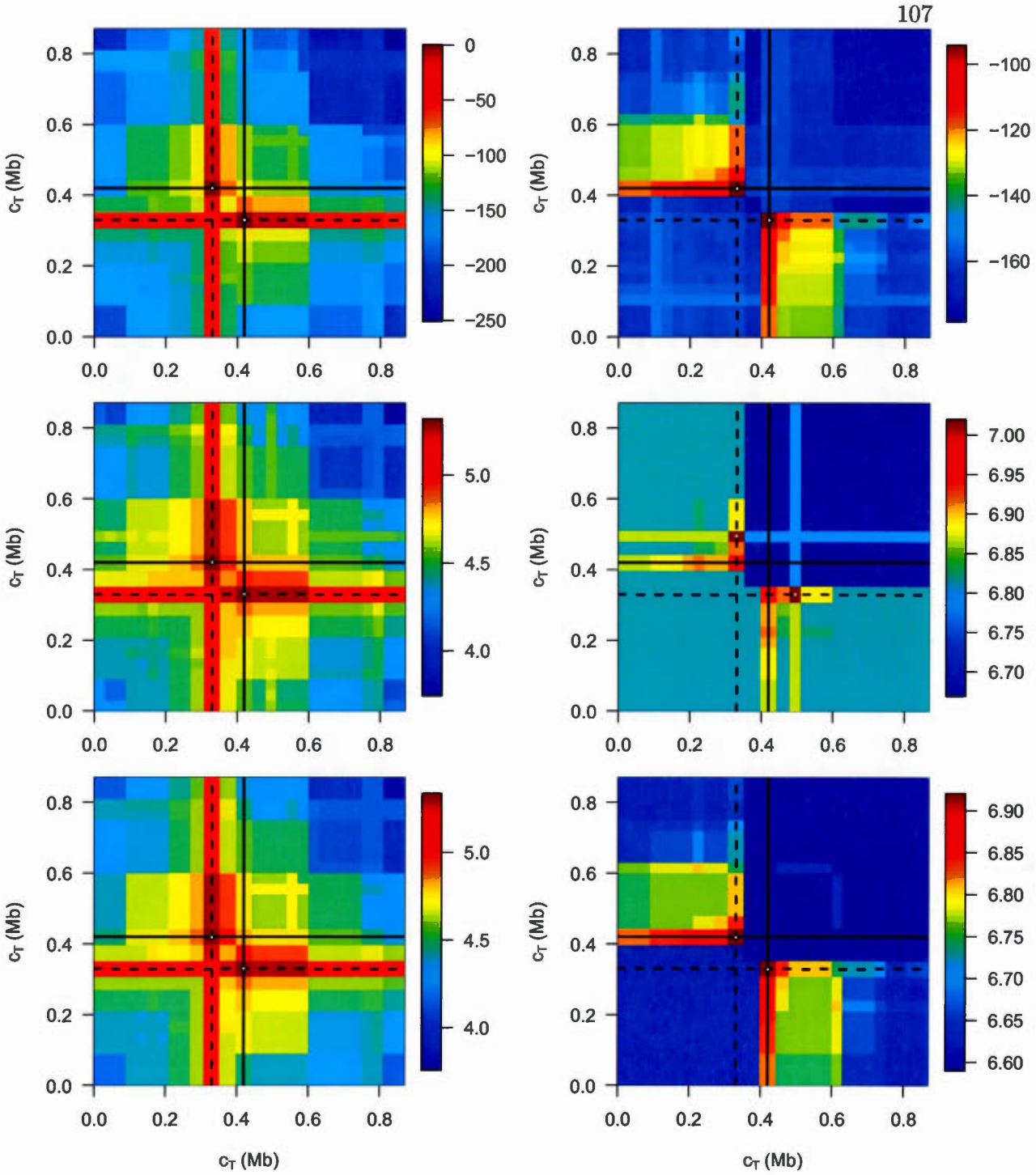

Figure 5.13: Modèle A (gauche) et modèle B (droite). 80 cas et 80 témoins. Estimation par la vraisemblance (haut), par le χ^2 (centre) et par le χ^2 de type 1 à gauche et de type 2 à droite (bas).

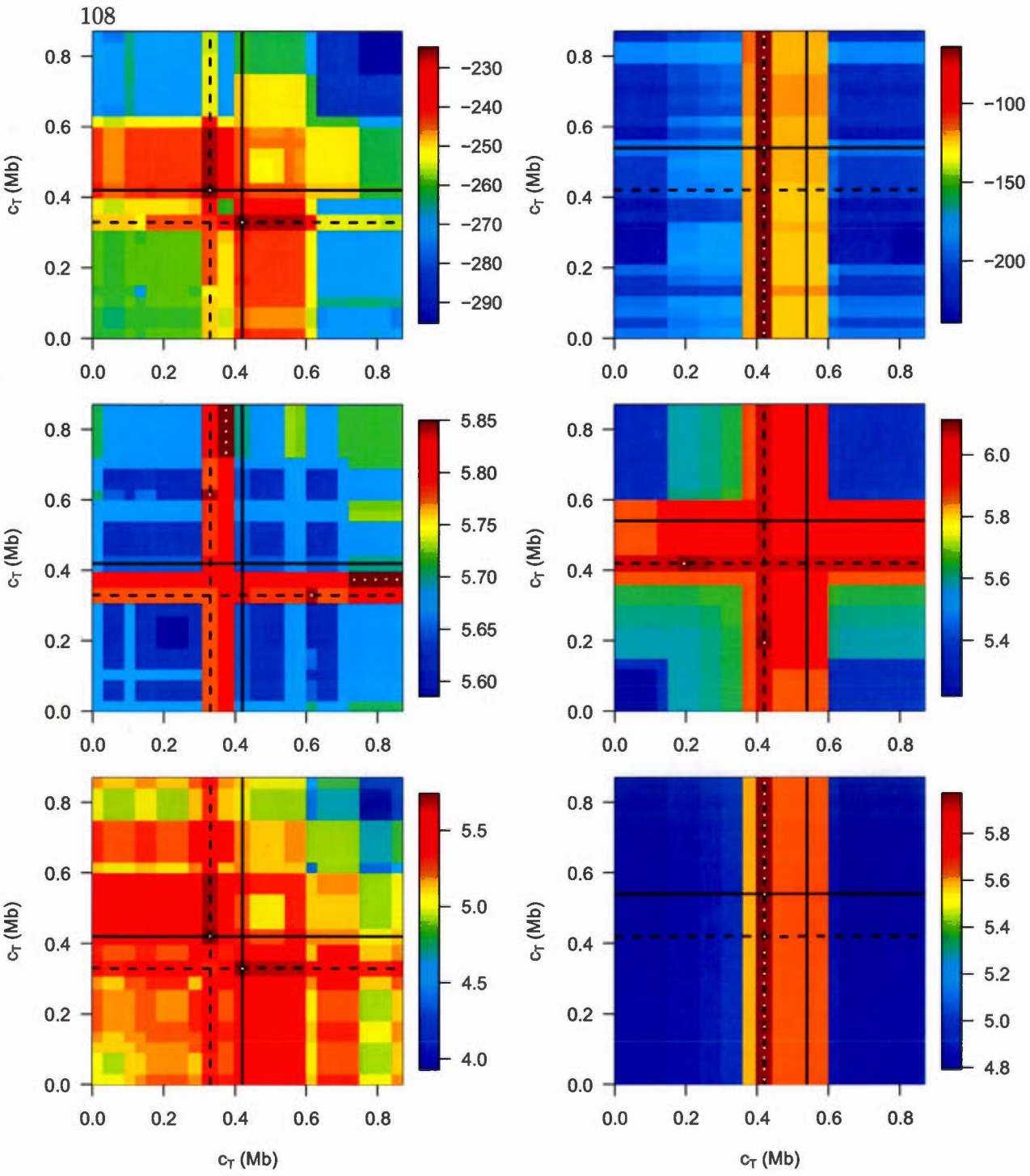

Figure 5.14: Modèle C (gauche) et modèle D (droite). 80 cas et 80 témoins. Estimation par la vraisemblance (haut), par le χ^2 (centre) et par le χ^2 de type 3 à gauche et de type 4 à droite (bas).

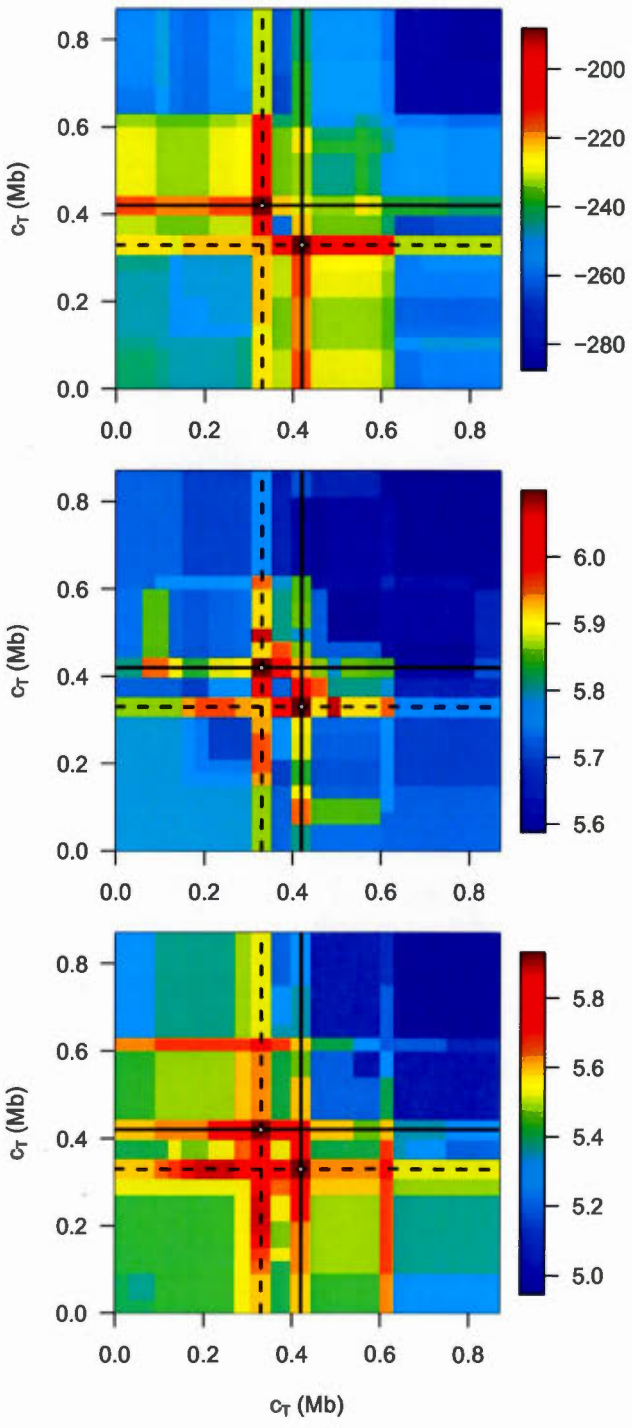

Figure 5.15: Modèle E. 80 cas et 80 témoins. Estimation par la vraisemblance (haut), par le χ^2 (centre) et par le χ^2 de type 5 (bas).

CONCLUSION

En définitive, nous avons développé et implémenté dans ce mémoire une façon de localiser deux mutations influençant un trait sur une séquence génétique en se basant sur la méthodologie *DMap* de Descary (2012) utilisant les graphes de recombinaison ancestraux. Cette méthode, *DMapInteraction*, repose sur la distribution proposée par Fearnhead et Donnelly (2001) pour générer plusieurs généalogies qui auraient pu être la source de nos données observées et utilise ensuite des estimateurs par le maximum de vraisemblance et des statistiques du χ^2 pour localiser les couples de SNPs qui ont le plus probablement donné lieu à un certain phénotype. Étant donné que ce sont maintenant des paires de TIMs qui sont recherchés, la complexité de notre algorithme est augmentée d'un facteur exponentiel par rapport à *DMap* et une partie importante de notre travail a été de viser à diminuer le temps d'exécution.

Les résultats obtenus avec des données simulées dans le chapitre 5 nous ont lancé sur plusieurs pistes de réflexion. On a vu, dans la section 5.3 où on utilise l'arbre généalogique de nos données, que l'utilisation d'un facteur correcteur estimant un échantillonnage aléatoire lorsqu'on emploie un échantillon stratifié, un échantillon plus facile à obtenir dans les études cliniques, nous permet d'améliorer nos détections. On a également remarqué que d'utiliser des tests du χ^2 tenant compte du type de phénotype auquel on est confronté nous offre des résultats intéressants tout en respectant mieux les conditions de ce type de test qu'un χ^2 plus général, étant donné les dimensions 2x2 du tableau. On peut finalement penser, en comparant les résultats obtenus à la section 5.2, où on simule des ARGs, avec la section 5.3, que cela nous prendrait un nombre de graphes beaucoup plus important pour

déterminer correctement la puissance de la capacité de détection de *DMapInteraction*, ce qui est impossible pour le moment à cause des temps excessifs nécessaires à la complétion d'un seul test. Néanmoins, étant donné que dans certains cas, même le χ^2 usuel, une méthode répandue en cartographique génétique, ne réussit à détecter les mutations causales, nous sommes conscients que la compréhension de certains types de maladie est encore à travailler et nous pensons que le fait de considérer la relation entre les individus dans notre méthode pourrait être un apport à la recherche sur ces maladies.

Il pourrait être intéressant pour améliorer la vitesse d'exécution de notre programme de trouver une façon d'introduire une vraisemblance composite à *DMapInteraction*. Ce type de vraisemblance décompose une vraisemblance usuelle en plusieurs, chacune calculée sur des fenêtres de marqueurs au lieu de la séquence complète. La vraisemblance composite a déjà été incorporée à *DMap* et a permis de réduire considérablement les temps de calcul, ce qui nous laisse croire que l'effet pourrait être le même dans *DMapInteraction*. On pourrait donc éventuellement avoir la possibilité de simuler plusieurs dizaines de milliers de graphes, comme on le souhaiterait, en un temps raisonnable.

RÉFÉRENCES

Brown, T. (2012). *Introduction to Genetics : A Molecular Approach.* Garland Science, Taylor & Francis Group.

Cordell, H. J. (2002). Epistasis : what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*(20), 2463–2468.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, *10*, 392–404.

Cordell, H. J. et Clayton, D. G. (2005). Genetic association studies. *Lancet*, *366*, 1121–1131.

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

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

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

Forest, M. (2010). *Cartographie génétique fine simultanée de deux gènes.* (Mémoire de maîtrise). Université du Québec à Montréal.

Hamzelou, J. (2016). Exclusive : World's first baby born with new "3 parent" technique. New Scientist. Récupéré de `https://www.newscientist.com/article/2107219-exclusive-worlds-first-baby-born-with-new-3-parent-technique/`

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

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

Lewontin, R. (1964). The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, *49*(1), 49–67.

Marchini, J., Donnelly, P. et Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, *37*(4), 413–417.

Marjoram, P. et Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genetics*, *7*(16), 1–9.

McVean, G. A. et Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B*, *360*, 1387–1393.

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

Morris, A. et Cardon, L. (2007). *Handbook of Statistical Genetics*, volume 1, chapitre Whole Genome Association, 1238–1263. John Wiley & Sons, (3 éd.).

Nordborg, M. (2007). *Handbook of Statistical Genetics*, volume 1, chapitre Coalescent Theory, 843–877. John Wiley & Sons, (3 éd.).

Phillips, P. C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, *9*(11), 855–867.

Slatkin, M. (2008). Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*, 477–485.

Wan, X., Yang, C., Yang, Q., Zhao, H. et Yu, W. (2013). The complete compositional epistasis detection in genome-wide association studies. *BMC Genetics*, *14*(7), 1–11.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276.

Ziegler, A. et König, I. R. (2006). *A Statistical Approach to Genetic Epidemiology*. Wiley-VCH.