Cartographie de gènes à caractères quantitatifs par déséquilibre de liaison

Simon Boitard

Université Paul Sabatier – Toulouse III

U.F.R. Mathématique Informatique Gestion

Cartographie de gènes à caractères quantitatifs par déséquilibre de Liaison

THÈSE

présentée et soutenue publiquement le 12 Décembre 2006

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ TOULOUSE III

Discipline : Mathématiques Appliquées
Option : Statistique

par

Simon BOITARD

Composition du jury

Rapporteurs :	M. Olivier FRANÇOIS
	Mme. Christine DILLMANN
Examinateurs :	Mme. Béatrice LAURENT-BONNEAU
	M. Alain CHARCOSSET
	M. Philippe BARET
	M. Claude CHEVALET
Directeurs de thèse :	Mme. Brigitte MANGIN
	M. Jean-Marc AZAïS

Unité de Biométrie et Intelligence Artificielle
Institut National de la Recherche Agronomique, BP 52627, 31326 Castanet-Tolosan cedex

Laboratoire de Statistique et Probabilités –U.M.R. 5583
Université Paul Sabatier, 31062 Toulouse cedex 9

Mis en page avec la classe thloria.

Remerciements

Un seul nom sur la couverture, et pourtant nombreux sont ceux qui ont aidé de près ou de loin à ce que ce travail aboutisse. Je saisis ici l'occasion de leur témoigner ma profonde reconnaissance.

J'aimerais tout d'abord remercier les deux rapporteurs, Christine et Olivier, pour tout le temps et l'intérêt qu'ils ont accordé à mon manuscrit et les perspectives qu'ils ont proposées. Je remercie également les examinateurs pour leur participation à la soutenance : eux aussi ont pris le temps de s'intéresser à mon travail.

Mes premières pensées vont ensuite vers mes directeurs, Brigitte et Jean-Marc. Toujours disponibles et prêts à m'écouter, ils ont su me donner des conseils précieux dans les moments importants et j'ai beaucoup appris à leur contact. Ces trois années furent passionnantes et très agréables.

Bien d'autres ont également contribué à l'avancée de cette thèse. Je pense notamment à Christine, Jihad, Patrice et Hubert que je ne reconnaîtrai décidément jamais sur son vélo. Un grand merci également à Lounès, mon meilleur conseiller d'orientation.

La difficulté d'une thèse n'est pas que d'ordre scientifique, mais également psychologique. Je peux dire qu'à ce titre j'ai été énormément aidé par l'accueil chaleureux de l'unité BIA dans son ensemble. Merci à Pascale, Jackie et Maïthé pour leur redoutable efficacité et à Martin, Abde et Mickael les pros de l'informatique. Une pensée toute spéciale pour les malheureux qui ont dû supporter ma mauvaise humeur et mes nombreux coups de téléphone : Sylvain l'intarissable blagueur, Nassolo la chanteuse, Mathias l'homme des carottes râpées, Nicklas le prof d'anglais et Olivier qui a beaucoup progressé depuis le début de sa thèse (en escalade bien sûr) !

Merci à tous les autres de la croisière pour les folles soirées passées ensemble : Céline, Erika, Arnaud, Carrére, Romain, Mélanie, David l'avignonais, Anne la parisienne, Laure la montpelliéraine, Gabriella, Ana, Iadine . . .décidément il y a vraiment de plus en plus de monde sur la croisière ! Et bien sûr merci à Kim-Anh de me supporter à chaque congrès . . .et accessoirement de m'avoir sauvé la vie à Alto Paraiso !

Mais il y a aussi une vie en dehors de l'INRA (si si !) et malgré tout ce que je viens de dire ces années toulousaines auraient peut-être été un calvaire sans Mehdi l'acrobate, Mathieu le père peinard, Jey le théoricien (minute), Julien le resto du coeur, Arnaud l'alter-moraliste, Vincent la force tranquille, Thierry et Chloé le couple des monologues, Raphael l'extorqueur de bouteilles à la cale sèche, Nico mon digne successeur, Serge qui aurait fait une apparition récente aux alentours de la paillote (mais personne ne s'en souvient vraiment) et Delphine qui ne saura peut-être jamais toute l'affection que j'ai

pour elle, hésitant éternellement entre lire cette page ou la suivante (ou peut-être celle d'encore après ?).

Je voudrais aussi saluer tous ceux que mon enclavement toulousain ne m'a pas permis de voir suffisamment ces dèrniéres années : Jean-Phi et Sab les crevards, Loïc l'homme des EDP, Gbour mon maître métabolique, Mathieu et ses déclarations pré-remplies, Xav qui ne connaîtra sans doute jamais Amandine, Tony et Champ presque morts noyés à Toulouse (mais pas pour les mêmes raisons), Antoine et Arnaud mes bientôt cheums, Greg qui m'aurait mis un A+ pour la thèse, Guillaume qui aurait contesté et Seb qui se serait endormi.

Et puis pour finir en beauté j'ai réservé le plus grand de tous les merci à mes petits parents, ma grand-mère et ma soeur qui m'ont toujours soutenu, pendant la thèse ainsi qu'au cours des nombreuses années qui l'ont précédée. Cette thèse leur est dédiée.

A mes parents,
ma grand-mère Suzanne
et ma soeur Magali

Table des matières

Introduction générale

Rechercher les liens entre le génotype d'un individu et les caractères physiques qu'il exprime est un des enjeux scientifiques fondamentaux de la biologie actuelle. Chez l'homme, la découverte de gènes liés à des maladies partiellement héréditaires est un atout considérable dans la lutte contre ces maladies. Dans le domaine de l'agriculture, la connaissance des gènes contrôlant certains caractères agronomiques d'intérêt (masse musculaire ou production laitière d'un animal, résistance d'une plante aux aléas climatiques ou aux maladies, ...) permet notamment de mieux sélectionner les espèces.

Au cours de ces dernières années, la biologie a connu une véritable révolution avec l'essor spectaculaire de la génomique, qui a notamment conduit au séquençage et à l'annotation du génome humain. Des travaux similaires sont également en cours pour plusieurs espèces animales telles que la souris, le cochon, le bœuf, le lapin .... Pour autant, la détection systématique de tous les gènes d'un génome ne donne pas d'information sur leur fonction. Cette information est apportée par d'autres méthodes. Par exemple, les méthodes de réseaux métaboliques tentent de comprendre les interactions entre protéines et autres molécules au niveau cellulaire. Les méthodes de **cartographie génétique** s'intéressent quant à elles à des fonctions plus globales. Ce sont des méthodes statistiques qui permettent, pour un caractère donné, d'identifier les zones du génome qui sont impliquées dans l'expression de ce caractère. Une fois qu'une zone suffisamment petite d'un chromosome a été identifiée, les données de séquençage pour cette zone peuvent alors être utiliséees et des analyses biologiques plus pointues permettent de détecter exactement la séquence d'ADN qui a une influence sur le caractère.

Le principe général des méthodes de cartographie génétique est d'observer différents individus de la même espèce et de confronter, pour chaque individu, le génotype avec la valeur du caractère étudié. Un traitement statistique est alors nécessaire pour estimer, à partir de ces informations, la position des gènes influant sur le caractère. La cartographie génétique existe depuis plusieurs décennies et était traditionnellement basée sur l'étude de familles d'individus au cours de plusieurs générations. Mais depuis une dizaine d'an-

nées, une nouvelle approche a été proposée. Cette approche est basée sur l'utilisation du **déséquilibre de liaison**, que l'on peut définir comme la corrélation, au niveau d'une population, entre les génotypes observés à diférentes positions du génome. Les méthodes de déséquilibre de liaison étudient des individus non apparentés. L'information portée par un tel échantillon est potentiellement plus importante que celle apportée par l'observation d'une famille, ce qui permet d'estimer plus précisément la position des gènes. A l'origine, cette approche a été proposée dans le contexte de la génétique humaine pour localiser des gènes de maladies. Ces maladies étaient modélisées comme des caractères discrets. L'utilisation du déséquilibre de liaison pour localiser des gènes influant sur des caractères quantitatifs (QTL) est beaucoup plus récente. L'objectif de ma thèse est de développer de telles méthodes.

Deux types de stratégies sont possibles pour cartographier, par déséquilibre de liaison, les gènes impliqués dans l'expression d'un caractère. La première consiste à utiliser, s'ils existent, les résultats d'études précédentes qui fournissent des zones du génome où il y a très certainement un ou plusieurs de ces gènes. L'utilisation du déséquilibre de liaison permet alors d'estimer plus précisément la position du ou des gènes dans chaque zone. Dans ce cas, on parle généralement de **cartographie fine** par déséquilibre de liaison. Les principaux résultats que j'ai obtenus dans ce domaine sont exposés dans les parties II et III du manuscrit.

L'autre stratégie consiste à analyser directement un très grand nombre de positions sur l'ensemble du génome et à tester, pour chaque position, si son génotype a une influence sur le caractère. Cette approche, très en vogue actuellement, est apparue à la faveur de récents progrès technologiques qui ont rendu possible le génotypage intensif des individus. Les méthodes utilisant ce type d'approche sont appelées des **études d'association**. Ma contribution à cette question est présentée dans la partie IV.

Un des objectifs de ce manuscrit est de donner une vision globale du domaine de la cartographie de gènes par déséquilibre de liaison. Pourquoi utiliser ce genre de stratégies pour cartographier des gènes ? Quels sont les problèmes statistiques à résoudre ? Quel genre de méthodes sont généralement employées ? Je tente de répondre à toutes ces questions dans la partie I du manuscrit. Cette partie est une compilation de résultats classiques de génétique mathématique, que j'ai réorganisés de manière à introduire progressivement les notions et les modèles nécessaires à la bonne compréhension du domaine.

Le chapitre 1 permet de comprendre ce qu'est la cartographie de gènes. J'y explique comment le génome d'une espèce peut être représenté par une carte munie d'une certaine

distance et comment l'influence des gènes sur un caractère peut être modélisée. Les notions d'haplotype et de taux de recombinaison, fondamentales dans la suite du manuscrit, y sont également définies.

Les méthodes de cartographie par déséquilibre de liaison sont en grande partie basées sur des modèles probabilistes de génétique des populations. Je présente donc ces modèles dans le chapitre 2. J'insiste particulièrement sur ceux de ces modèles qui permettent de représenter l'évolution des fréquences d'allèles ou des fréquences d'haplotypes au sein d'une population, car ils seront au cœur des travaux présentés dans la suite du manuscrit.

La notion centrale de déséquilibe de liaison est l'objet du chapitre 3. Après avoir défini cette notion, je re-démontre un résultat classique qui illustre pourquoi cette notion est importante pour la cartographie de gènes. J'essaie également, en m'appuyant sur divers résultats connus, de lever certaines ambigüités au sujet de la loi stationnaire du déséquilibre de liaison. Pour compléter cette partie, un article de revue auquel j'ai contribué est fourni en annexe du manuscrit.

Je donne finalement au chapitre 4 une description précise du cadre statistique de la cartographie fine de gènes par déséquilibre de liaison et des hypothèses qui seront faites dans les partie suivantes. Je décris ensuite les différentes méthodes d'estimation existantes. Mon travail se situe plus particulièrement dans celle de ces approches que l'on nomme généralement **approche fréquentiste**. Dans cette approche, la vraisemblance de la position du gène recherché est calculée en intégrant par rapport à la distribution de probabilités des fréquences d'haplotypes dans la population. Les travaux présentés dans les parties II et III permettent de calculer la vraisemblance de manière plus précise que cela n'est fait actuellement dans la litérature.

Dans les méthodes fréquentistes, la vraisemblance de la position du gène est généralement approchée par un développement limité d'ordre 1 (exceptionnellement 2), car la distribution de probabilité des fréquences d'haplotypes pour des modèles à plusieurs loci avec recombinaison est mal connue. L'idée directrice de la partie II est de remédier à ce problème en décrivant plus précisément cette distribution.

Dans le chapitre 5, je m'intéresse donc à la distribution de probabilité des fréquences d'haplotypes pour un modèle de diffusion à deux loci bialléliques avec recombinaison. Le générateur de cette diffusion est connu depuis longtemps. Pourtant, contrairement au modèle à un locus (ou de manière équivalente à plusieurs loci sans recombinaison), l'expression de la densité n'a jamais été trouvée et seuls certains moments de la distribution ont pu être obtenus sous des hypothèses particulières. C'est donc un problème difficile et dont la portée dépasse largement le cadre de la cartographie de gènes. Je propose dans ce

chapitre une méthode numérique approchée, basée sur les différences finies, pour résoudre les équations de Kolmogorov associées à ce processus de diffusion. Appliquée à l'équation projective de Kolmogorov, cette méthode me permet de calculer la densité de transition $f(y^0, y, \tau)$ du processus, pour un temps τ et un vecteur de fréquences initiales y^0 fixés, et pour tous les vecteurs de fréquences finales y tels qu'il n'y a eu fixation à aucun des 2 loci. Pour tester la précision de la solution ainsi obtenue, j'étudie certaines propriétés de la distribution pour lesquelles des résultats théoriques ont été démontrés. Ceci me permet d'identifier les valeurs des paramètres pour lesquelles la méthode est efficace et les valeurs pour lesquelles elle l'est moins. Je montre également que la complexité algorithmique de cette méthode est inférieure à celle d'autres méthodes classiques auxquelles on pourrait penser pour résoudre le même problème. Tout ce travail est présenté sous la forme d'un article qui a été soumis en juin 2006 à la revue *Theoretical population biology.*

Dans le chapitre 6, j'étudie une application de la méthode numérique du chapitre 5 au problème de la cartographie fine par déséquilibre de liaison. L'objectif est d'estimer le taux de recombinaison c entre deux loci bialléliques dans le cas où on observe, parmi un échantillon de $2N_s$ haplotypes, les effectifs respectifs des 4 haplotypes possibles. Si un de ces loci est un marqueur de position connue et l'autre un gène d'intérêt de position inconnue, on voit qu'il s'agit bien d'un problème de cartographie. C'est cependant un problème un peu idéal car on suppose connu le génotype du gène à positionner, alors qu'en pratique on ne dispose que d'un phénotype lié à ce gène. Je propose dans ce chapitre trois estimateurs de maximum de vraisemblance pour c, où la vraisemblance est intégrée par rapport à la distribution de probabilité des fréquences d'haplotyes aux 2 loci dans la population. La première méthode simule cette distribution et calcule la vraisemblance par Monte Carlo. La seconde utilise une approximation d'ordre 1 de la vraisemblance et utilise simplement les espérances des fréquences d'haplotypes. La troisième enfin utilise l'algorithme présenté au chapitre 5 pour résoudre l'équation rétrograde de Kolmogorov vérifiée par la vraisemblance. Je montre sur différents exemples que cette dernière méthode donne des estimations d'une bonne précision pour un temps de calcul qui reste raisonnable. C'est par conséquent une voie à explorer.

Le chapitre 7 peut être vu comme un complément du chapitre 5. J'y propose plusieurs pistes pour améliorer la précision de la méthode numérique. Ces pistes sont principalement liées à une meilleure prise en compte des conditions aux bords associées aux équations de Kolmogorov. J'évoque également différents types de problèmes pour lesquels il serait sans doute intéressant d'utiliser cette méthode. Ce chapitre ouvre donc diverses perspectives pour poursuivre le travail de la partie II.

Une autre voie à explorer pour améliorer la précision d'estimation des méthodes de cartographie fréquentistes est l'augmentation du nombre de marqueurs qui sont utilisés simultanément pour calculer la vraisemblance en une position de la carte. Je présente dans la partie III une méthode de cartographie fine de QTL qui va dans ce sens.

Cette méthode a fait l'objet d'une publication d'équipe dans la revue *BMC Genomics*. Cette publication constitue le chapitre 8. La méthode que nous avons développée permet d'estimer par maximum de vraisemblance la position x d'un QTL sur une carte génétique. Pour chaque position x, on utilise l'information des deux marqueurs flanquants et une approximation d'ordre 1 de la vraisemblance est calculée à l'aide de cette information. Une méthode similaire avait déjà été développée dans l'équipe de recherche INRA au sein de laquelle j'ai effectué ma thèse, mais cette méthode ne pouvait tenir compte que d'un marqueur à la fois pour calculer la vraisemblance. Pour utiliser plusieurs marqueurs, on faisait le produit des vraisemblances pour chaque marqueur en supposant leur indépendance. Pour permettre à la méthode de tenir compte des corrélations entre les vraisemblances des marqueurs proches, j'ai dérivé une expression de l'espérance des fréquences d'haplotypes sous un modèle de Wright-Fisher à 3 loci avec recombinaison. Les 3 loci représentent le QTL et les 2 marqueurs flanquants. Pour faire ce calcul, j'ai supposé toutefois que les fréquences alléliques aux marqueurs étaient constantes et en équilibre. Ce résultat a été intégré dans le programme existant. Sur plusieurs scénarios de simulations, nous avons observé que la nouvelle méthode obtenue, baptisée HAPim, était effectivement plus précise que l'ancienne. HAPim a également fourni des résultats aussi précis que d'autres méthodes multi-marqueurs basées sur la notion d'identité par descendance. Ces méthodes nécessitent cependant un temps de calcul bien plus important.

Le chapitre 9 apporte quelques compléments à propos des évolutions possibles de HAPHim. Je reviens sur les hypothèses faites lors du calcul de l'espérance des fréquences d'haplotypes, et explique comment étendre ces calculs pour intégrer plus de marqueurs simultanément ou bien tenir compte de la mutation ou de la sélection.

La partie IV est composée du seul chapitre 10, mais j'ai souhaité le distinguer des autres dans la mesure où il traite des études d'associations et non de la cartographie fine de gènes. Je me suis interessé à ces méthodes à l'occasion d'un projet de détection de QTL pour la qualité de la viande bovine du nom de GeMqual. Dans le chapitre 10, je commence par présenter les méthodes d'association en général, et j'évoque plus particulièrement les problèmes liés à la présence d'une structure dans la population étudiée. Cette problématique se retrouve dans le projet GeMqual, où les animaux sont issues de différentes races connues. Je présente ensuite une étude que j'ai réalisée dans le cadre du

projet GeMqual, et dont le but est d'évaluer les propiétés asymptotiques d'un test de type Transmission Desequilibrium Test (TDT) quantitatif dans le cas d'une population structurée. J'utilise un modèle linéaire de régression des phénotypes par les génotypes qui avait déjà été proposé dans la litérature. Ce modèle suppose que les génotypes sont des variables aléatoires dont la loi dépend de la sous-population, et que leurs effets sur les phénotypes sont également différents dans chaque population. Dans le cadre de ce modèle, je montre que le TDT, contrairement à ce qui est souvent avancé, est légèrement libéral. Je compare aussi la puissance de détection du TDT à celle d'un test qui tient compte de la structure de la population. Ce travail a été intégré au rapport final du projet GeMqual, mais quelques précisions ont été apportées depuis et ont donc été intégrées au chapitre 10.

Première partie

Cadre d'étude

Chapitre 1

Gènes et caractères : définitions et modèles

L'objectif central de ce premier chapitre est de définir ce qu'est la cartographie fine de gènes. Il introduit pour cela les notions fondamentales de distance génétique et de carte génétique, ainsi que les modèles mathématiques décrivant la relation entre identité physique et identité génétique d'un individu. Les notions de génétique élémentaires utiles à la compréhension des différents modèles sont également rappelées.

1.1 Le génome

L'identité biologique de chaque être vivant est contenue dans la molécule d'ADN. Celle-ci s'organise en plusieurs *chromosomes* dont le nombre diffère d'une espèce à une autre. Sur chaque chromosome, il existe des zones, appelées *gènes*, jouant un rôle dans l'expression d'un caractère particulier (par exemple la couleur des yeux ou le groupe sanguin chez l'homme) ; chaque gène existe sous différentes versions appelées *allèles*. Tous les individus d'une même espèce possèdent les mêmes gènes, qui constituent le *génome* de cette espèce. Mais chacun possède un ensemble d'allèles qui lui est propre et que l'on nomme *génotype*. Chez les espèces diploïdes et à reproduction sexuée telles que la plupart des animaux ou des végétaux, les chromosomes sont associés par paires : ces chromosomes *homologues*, l'un provenant de la mère et l'autre du père, possèdent les mêmes gènes mais avec des allèles a priori différents, si bien que l'expression d'un caractère dépend généralement de la combinaison des deux allèles pour un gène donné. Nous reviendrons sur ce point dans la section 1.4.

Pour traduire ces phénomènes biologiques, nous utilisons le modèle mathématique suivant. Chaque chromosome est représenté par un intervalle réel. L'emplacement d'un

gène sur un chromosome, appelé son *locus*, est alors vu comme un point sur cet intervalle (la taille des gènes étant en principe négligeable devant la taille totale du chromosome). Les différents allèles d'un locus sont représentés par des étiquettes, généralement des entiers ($i = 1, \ldots, I$, I nombre total d'allèles) ou des lettres (A, B, C ...). Le génotype d'un individu diploïde au niveau d'un locus est représenté par une paire non ordonnée d'allèles que l'on note i_1/i_2. On dit que l'individu est *homozygote* pour ce locus si $i_1 = i_2$ et *hétérozygote* pour ce locus si $i_1 \neq i_2$.

Pour l'étude simultanée de plusieurs loci, on appelle *haplotype* toute suite ordonnée d'allèles formée en prenant un allèle de chaque locus. Nous adopterons ici la notation $j = (i_1, i_2, \ldots, i_L)$ où L est le nombre total de loci et i_l l'indice de l'allèle au locus l. Au niveau de ces L loci, l'identité génétique d'un individu diploïde est donc caractérisée par la donnée de deux haplotypes $j_1 = (i_1^{(1)}, \ldots, i_L^{(1)})$ et $j_2 = (i_1^{(2)}, \ldots, i_L^{(2)})$. L'un de ces haplotypes est hérité du père et l'autre est hérité de la mère. La paire non ordonnée formée par ces deux haplotypes est appellée son *diplotype*, et est notée j_1/j_2. Cependant cette information n'est pas toujours disponible, et on utilise donc souvent la notion de génotype pour L loci, qui est une extension de la notion de génotype pour un locus. C'est en effet la suite ordonnée des génotypes à chaque locus, que l'on note $k = (i_1^{(1)}/i_1^{(2)}, \ldots, i_L^{(1)}/i_L^{(2)})$.

Dans la section suivante, nous allons compléter ce modèle par la définition d'une distance sur les chromosomes.

1.2 Crossing-over et distance génétique

1.2.1 Crossing-over

Toutes les cellules d'un individu contiennent une copie de son génotype. Nous avons vu que chez les espèces diploïdes à reproduction sexuée, cette information était constituée de n paires de chromosomes, soit $2n$ chromosomes. Les cellules sexuelles ou *gamètes* constituent cependant une exception car elles ne contiennent qu'un chromosome de chaque paire. Ainsi au moment de la reproduction, un gamète du père et un gamète de la mère fusionnent pour donner à nouveau une cellule avec des paires de chromosomes, qui sera à l'origine de toutes les cellules du nouvel individu.

La production des gamètes est assurée par un processus de division cellulaire appelé *méiose* qui crée quatre gamètes avec n chromosomes à partir d'une cellule normale contenant $2n$ chromosomes. Mais le plus important est que ces quatre gamètes ne contiennent pas forcément un des deux chromosomes de chaque paire, mais souvent un mélange des deux. En effet, au cours de la méiose, les chromosomes homologues s'apparient : il arrive

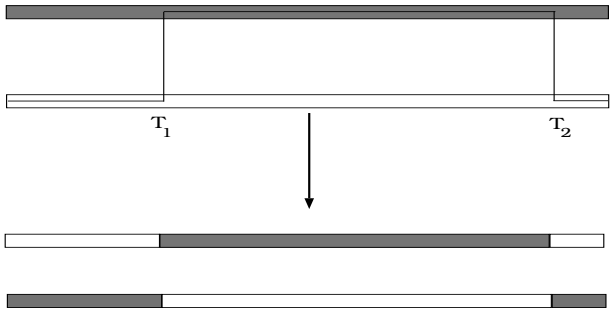


FIG. 1.1 – Echange de matériel génétique entre deux chromosomes homologues lors de la méiose. Deux crossing-over ont lieu en T_1 et T_2

alors qu'ils se cassent en plusieurs endroits et échangent leur matériel génétique. Ce phénomène se produit de façon aléatoire et se nomme le *crossing-over* (voir figure 1.1). Il est essentiel pour la diversité d'une espèce car il permet la création de nouvelles combinaisons alléliques.

Prenons en effet l'exemple d'un individu qui, pour deux loci donnés situés sur le même chromosome, possède l'haplotype $(1, 1)$ hérité de sa mère et l'haplotype $(2, 2)$ hérité de son père. Si, au cours de la méiose, il n'y a pas de crossing-over entre les deux loci, ou qu'il y en a un nombre pair, les gamètes formés seront également de type $(1, 1)$ ou $(2, 2)$. Mais s'il y a un nombre impair de crossing-over entre les loci, alors les gamètes formés seront de type $(1, 2)$ ou $(2, 1)$, qui sont des combinaisons nouvelles par rapport au parent.

On appelle *taux de recombinaison* entre deux loci d'un même chromosome, et on note c, la probabilité qu'il y ait un nombre impair de crossing-over entre ces loci au cours de la méiose. On suppose que cette quantité est la même pour tous les individus d'une même espèce et qu'elle est constante au cours du temps. On dit que deux gènes sont *liés* si $c < 1/2$ et *indépendants* si $c = 1/2$, puisque si $c = 1/2$ il n'y a aucune corrélation entre l'allèle transmis au premier locus et l'allèle transmis au second locus. Pour deux loci issus de chromosomes distincts, la notion de crossing-over intervenant entre les loci n'a pas de sens. Par convention on prend donc $c = 1/2$ car ces loci sont a priori indépendants.

1.2.2 Distance génétique

Deux loci qui sont physiquement proches ont une probabilité plus faible de recombiner. Le taux de recombinaison semble donc pouvoir fournir une bonne mesure de la distance entre loci. Mais cette distance n'est pas additive, car l'effet des crossing-over s'annule quand leur nombre est pair. On définit donc plutôt la *distance génétique* comme une fonction du taux de recombinaison, de façon à restaurer la propriété d'additivité.

Le choix le plus classique est celui de la *distance de Haldane* (Haldane, 1919), définie par

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

Elle s'obtient en supposant que les crossing-over sont indépendants et que leur nombre suit un processus de Poisson d'intensité 1. Quand c tend vers 0, on peut constater que cette distance est équivalente à c. Nous supposerons d'ailleurs souvent que $d \approx c$ dans la suite de ce mémoire, car nous travaillerons avec des loci très proches. La distance génétique s'exprime en *centiMorgans* (cM) : 1 cM correspond approximativement à 1% de recombinaison entre les loci.

Remarque 1 *Il pourrait sembler naturel de choisir une définition plus physique de la distance. En effet chaque chromosome est composé d'une longue chaîne de molécules appelées bases, et le nombre de bases d'ADN entre deux loci paraît donc un choix logique. Mais en réalité cette définition est peu pratique pour nous car à l'échelle où on se place elle est très difficile à mesurer expérimentalement. Une distance basée sur la recombinaison demande beaucoup moins de moyens techniques et est en outre plus adaptée à des applications statistiques. Dans les cas ou l'on travaille sur de très petites distances physiques (typiquement à l'intérieur d'un gène), on utilise souvent des règles d'équivalence. Chez l'humain par exemple, on estime que* 10^6 *bases = 1cM. Mais ce type d'équivalences a été fortement discuté ces dernières années, depuis que des travaux ont mis en évidence l'existence de* hot spots*, zones du génome où le taux de recombinaison serait beaucoup plus élevé qu'en moyenne (Jeffreys et al., 2001; Gabriel et al., 2002).*

1.3 Marqueurs et cartes génétiques

Un *marqueur génétique* est un caractère contrôlé par un locus unique et dont l'expression n'est pas influencée par l'environnement. Pour un individu donné, l'observation de ce caractère doit permettre de déterminer sans ambiguïté le génotype au locus correspondant. Par le passé, les seuls marqueurs disponibles étaient des marqueurs morphologiques, tels que la couleur des yeux ou autre. Mails ils ont été progressivement abandonnés car ils sont en général contrôlés par plusieurs loci et car leur expression est partiellement influencée par l'environnement. Aujourd'hui, on utilise donc plutôt des marqueurs moléculaires ou biochimiques. Au cours de la dernière décennie, de très nombreux progrès ont été réalisés dans ce domaine.

Il existe de nombreux types de marqueurs. Deux d'entre eux sont particulièrement uti-

lisés aujourd'hui : les *single nucleotid polymorphisms* (SNP) et les *microsatellites* (MST). Les SNP sont des loci constitués d'une unique base d'ADN au niveau de laquelle le taux de mutation est très faible. Ils sont donc le plus souvent *bialléliques* (il n'existe que deux allèles distincts). Les MST sont des séquences répétées de quelques bases d'ADN. Ils sont *multialléliques* (il peut exister de nombreux allèles distincts). Les génotypes des SNP et MST sont directement observables par des expériences biologiques. Ces marqueurs possèdent en outre l'énorme avantage d'être présents en très grand nombre et répartis tout le long des génomes des mammifères (environ cent mille MST et trois millions de SNP chez l'homme).

La *carte génétique* d'une espèce est une représentation du génome de cette espèce. Elle contient les marqueurs que l'on a pu identifier sur chaque chromosome, ainsi que leur ordre relatif et la distance génétique qui les sépare. Les marqueurs jouent donc le rôle de "balises", de "graduations" sur ces cartes. Pour estimer les taux de recombinaison entre marqueurs, et donc les distances génétiques, l'idée naturelle est d'observer le génotype d'un individu ainsi que celui de tous les gamètes qu'il produit. C'est ce que l'on fait indirectement en observant les génotypes de couples de parents ainsi que ceux de leurs enfants. Les procédures exactes de construction de cartes sont de simples routines aujourd'hui, gérées par des logiciels rodés tels que MapMaker (Lander et al., 1987), JoinMap (Stam, 1993) ou CartaGène (Givry et al., 2005). Nous ne rentrerons pas dans les détails des différentes étapes de ces procédures. Sachons simplement que pour l'homme ou la plupart des espèces animales et végétales étudiées en agronomie, de telles cartes génétiques existent, et que pour les régions du génome les plus intéressantes on dispose souvent de marqueurs distants de moins de 1cM.

Remarque 2 *Si, pour deux loci donnés, un individu possède les haplotypes* $(1,1)$ *et* $(1,2)$*, les gamètes issus d'une recombinaison entre ces deux loci sont également de type* $(1,1)$ *ou* $(1,2)$*, et ne peuvent donc pas être distingués des gamètes pour lesquels il n'y a pas eu de recombinaison. On parle alors de recombinaisons non visibles. Ce type de recombinaisons doit également être pris en compte dans les méthodes d'estimation du taux de recombinaison.*

1.4 Modélisation des caractères

Un caractère dont la valeur est déterminée entièrement par le génotype d'un seul locus est dit *mendélien*. A une combinaison allélique de ce locus correspond exactement une valeur du caractère, et le caractère est donc discret. L'exemple typique est celui

de certaines maladies génétiques comme la maladie d'Huntington (Snell et al., 1989), l'épilepsie (Lehesjoki et al., 1993) ..., où une mutation génétique au niveau d'un locus particulier a créé un allèle responsable du dysfonctionnement. On note souvent d cet alllèle, pour "diseased". Par exemple pour une maladie récessive, les individus de genotype d/d sont malades alors que les individus ayant au moins un allèle différent de d sont sains.

Ce type de déterminisme est très rare dans la nature. En général, les valeurs prises par un caractère, aussi appelées *phénotypes*, sont plutôt représentées comme des variables continues dépendant d'un très grand nombre de loci ainsi que de conditions externes liées à l'environnement des individus. Pour modéliser l'influence d'un groupe de loci sur un caractère, on suppose donc que, pour chaque individu, la valeur du caractère est une variable aléatoire continue Z dont la densité ψ_k dépend du génotype k aux loci considérés. On obtient ainsi le modèle de mélange

$$Z = \sum_{k=1}^{K} \mathbb{1}_{G=k} Z_k$$

où K est le nombre total de génotypes possibles pour les loci considérés, G représente le génotype et $\mathbb{1}_{G=k}$ est la variable indicatrice valant 1 si $G = k$ et 0 sinon. Chaque variable Z_k a une distribution ψ_k.

Différents choix de modélisation sont ensuite possibles. Dans la grande majorité des travaux, on suppose que conditionellement à la conaissance des génotypes, les phénotypes sont des variables de loi normale de variance constante σ^2 et de moyenne m_k dépendant du génotype, ce qui conduit au modèle linéaire

$$Z_k = m_k + \epsilon$$

où m_k est l'effet du génotype pour l'ensemble des loci considérés et ϵ est une variable gaussienne d'espérance nulle représentant les autres effets génétiques et environnementaux. On suppose généralement que les effets des loci sont additifs. Avec la notation $k = (i_1^{(1)}/i_1^{(2)}, \ldots, i_L^{(1)}/i_L^{(2)})$ introduite en 1.1, ceci donne le modèle

$$Z_k = \sum_{l=1}^{L} m_{i_l^{(1)}/i_l^{(2)}} + \epsilon$$

où $m_{i_l^{(1)}/i_l^{(2)}}$ est l'effet du génotype $i_l^{(1)}/i_l^{(2)}$ du locus l et L est le nombre total de loci considérés. Signalons cependant que l'interaction entre plusieurs gènes, aussi appelée *épistasie*, est considérée aujourd'hui comme un phénomène important en génétique. Plusieurs tra-

vaux récents ont donc étudié des modèles incluant ce genre d'interactions (Xu, 2003; Marchini et al., 2005).

Pour modéliser l'effet d'un locus l sur le phénotype, nous suivrons l'approche classique décrite dans (Falconer et Mackay, 1996) : le locus est supposé biallélique d'allèles Q et q et $m_{i_l^{(1)}/i_l^{(2)}}$ est paramétré de la façon suivante :

$$m_{i_l^{(1)}/i_l^{(2)}} = \begin{cases} \mu_l + a_l & \text{si} \quad i_l^{(1)}/i_l^{(2)} = Q/Q \\ \mu_l + d_l & \text{si} \quad i_l^{(1)}/i_l^{(2)} = Q/q \\ \mu_l - a_l & \text{si} \quad i_l^{(1)}/i_l^{(2)} = q/q \end{cases} \tag{1.1}$$

où μ_l est l'effet moyen du locus, a_l son effet additif et d_l son effet de *dominance*. Si $d_l = 0$, on a un modèle purement additif qui équivaut à un modèle de regression sur le nombre d'allèles Q dans le génotype du locus (0, 1, ou 2). Si c'est le cas pour tous les loci considérés, et qu'il n'y a pas d'interactions entre loci, on peut écrire le modèle général à L loci comme un modéle d'haplotypes additif (à condition évidemment que les haplotypes soient connus). Avec la notation $k = j_1/j_2$, ceci donne

$$Z_k = m_{j_1} + m_{j_2} + \epsilon$$

Certains loci ont des effets plus importants que les autres. Si la distribution des phénotypes observés pour un échantillon d'individus est bimodale, on dit qu'on est en présence d'un *gène majeur*. Le cas des individus malades ou non malades constitue un exemple extrême. Dans le cas contraire on parle de *quantitative trait locus* (QTL). Pour détecter un QTL, il est nécessaire d'utiliser en plus des phénotypes l'information provenant de marqueurs.

1.5 Cartographie génétique : approches classiques

La *cartographie génétique* consiste à rechercher des gènes ou des groupes de gènes ayant une influence sur un caractère donné. On parle de cartographie de QTL quand le caractère en question est continu. Considérant une zone plus ou moins étendue de la carte génétique d'une espèce, on se pose généralement deux questions :

1. Y a-t-il dans cette zone un ou plusieurs QTL pour le caractère considéré ? En termes statistiques, cette question se formule comme un test d'hypothèse de type H_0 : "Pour tous les groupes de loci de la zone, la distribution ψ_k des phénotypes est la même quel que soit le génotype k de ce groupe" contre H_1 : "Il existe au moins un groupe

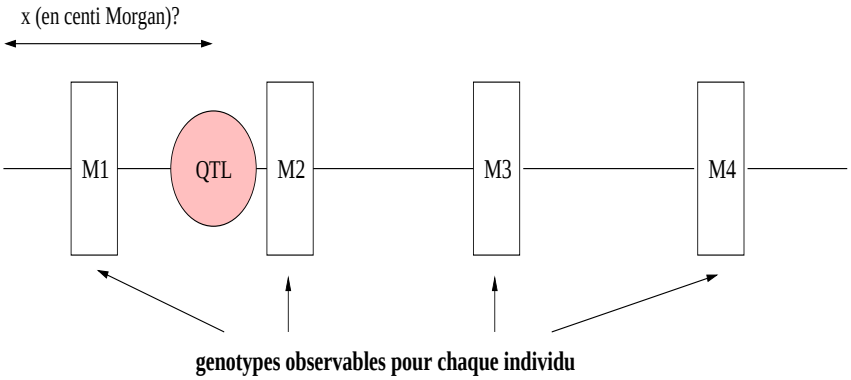


FIG. 1.2 – Estimation de la position d'un QTL sur une carte de marqueurs

de loci et deux génotypes k_1 et k_2 de ce groupe tels que $\psi_{k_1} \neq \psi_{k_2}$".

2. A supposer qu'il y ait des QTL dans la zone, quelle est leur position la plus probable ? Il s'agit donc ici d'un problème d'estimation de paramètres, en l'occurrence les positions des QTL sur la carte génétique (voir Figure 1.2).

Pour répondre à ces deux questions, nous disposons d'un échantillon de N_s individus de la même espèce et nous observons leurs phénotypes z_n, $n = 1, \ldots, N_s$, et leurs génotypes m_n, $n = 1, \ldots, N_s$ pour les L marqueurs de la zone. Pour autant les QTL peuvent tout à fait se trouver entre deux marqueurs, et leurs génotypes sont alors inconnus.

Dans l'approche classique appelée *analyse de liaison*, les individus considérés forment des familles de plusieurs générations dont le pedigree est connu. L'idée est alors de retracer la transmission des allèles des QTL au cours des générations et donc en quelque sorte de compter le nombre de recombinaisons par rapport aux différents marqueurs de la carte. Cette approche est en particulier très pratique pour des populations de plantes ou d'animaux dans lesquelles on peut croiser les individus suivant des schémas spécifiques (*backcross, lignées recombinantes* ...) de manière à simplifier les inférences sur la transmission des allèles. Mais sur des zones de faible distance génétique, la probabilité d'observer une recombinaison au cours des quelques générations dont sont issus les individus observés devient très petite. Ainsi il semble difficile d'obtenir par l'analyse de liaison des intervalles de confiance inférieurs à 5cM pour la position d'un QTL (Bodmer, 1986; Boehnke, 1994).

Les parties II et III de ce manuscrit sont consacrées au problème de la *cartographie fine* de QTL : on suppose que des méthodes telles que l'analyse de liaison ont déjà permis d'identifier une zone où il y a **exactement un QTL**, et on cherche à obtenir une estimation plus précise de sa position. Cet objectif s'inscrit donc dans la deuxième des questions

définies ci-dessus. Le problème traité dans la partie IV relève en revanche de la première question, puisqu'il s'agit de comparer différentes statistiques permettant de tester si un locus est ou n'est pas un QTL.

Pour améliorer la puissance de détection d'un QTL (question 1) ou la précision d'estimation de sa position (question 2), il faut être capable d'utiliser les recombinaisons intervenues depuis de très nombreuses générations. La notion de *déséquilibre de liaison*, présentée dans le chapitre 3, peut être exploitée dans ce sens.

Références

Bodmer, W. (1986). Human genetics : the molecular challenge. *Cold Sring Harbor Symp Quant Biol*, *51*, 1-13.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies : implication for the positional cloning of human disease genes. *Am J Hum Genet*, *55*, 379-390.

Falconer, D., et Mackay, T. (1996). *Introduction to quantitative genetics* (4th ed.). Longman.

Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., et al. (2002). The strucuture of haplotype blocks in the human genome. *Science*, *296*, 2225-2229.

Givry, S. de, Bouchez, P., Chabrier, P., Milan, D., et Schiex, T. (2005). Carthagene : multipopulation integrated genetic and radiated hybrid mapping. *Bioinformatics*, *21(8)*, 1703-1704.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics*, *8*, 299-309.

Jeffreys, A., Kauppi, L., et Neumann, R. (2001). Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nat. Genet.*, *29*, 217-222.

Lander, E., Green, P., Abrahamson, J., Barlow, A., Daly, M., Lincoln, S., et al. (1987). Mapmaker : an iterative computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, *1*, 174-181.

Lehesjoki, A., Koskiniemi, J., Norio, R., Tirrito, S., Sistonen, P., Lander, E., et al. (1993). Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21 : linkage disequilibrium allows high resolution mapping. *Hum. Mol. Genet.*, *2*, 1229-1234.

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

Snell, R., Lazarou, L., Youngman, S., Quarrell, O., Wasmuth, J., Shaw, D., et al. (1989).

Linkage disequilibrium in huntington's disease : an improved localization for the gene. *J Med Genet*, *26*, 673-675.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package : Joinmap. *The Plant Journal*, *3*, 739-744.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics*, *163*, 789-801.

Chapitre 2

Quelques modèles de représentation des populations en génétique

Les génotypes que nous observons aujourd'hui dans les échantillons d'individus utilisés pour la cartographie génétique sont le produit d'une longue suite d'événements (mutations, migrations, recombinaisons ...) intervenus au cours des générations passées. Reconstruire -dans une certaine mesure- l'histoire de ces événements est le moyen utilisé par les méthodes de cartographie dites de déséquilibre de liaison pour estimer la position des QTL. Avant de décrire précisément ces méthodes aux chapitres 3 et 4, nous présentons ici quelques modèles classiques permettant de représenter l'évolution du matériel génétique au niveau d'une population, et sur lesquels sont basés les développements des chapitres suivants. Les modèles de la section 2.1 concernent les fréquences d'haplotypes dans une population. Ceux de la section 2.2 décrivent la généalogie d'un échantillon de chromosomes.

2.1 Évolution des fréquences d'haplotypes

En génétique, on représente souvent une population par les fréquences des allèles ou des haplotypes portés par les individus de cette population. Différents modèles permettent de décrire l'évolution de ces fréquences. Je commencerai par présenter les modèles à un seul locus dans une perspective historique. Pour plus de détails on pourra se reporter à (Tavaré et Zeitouni, 2004). Je décrirai ensuite les modèles à plusieurs loci.

2.1.1 Modèles à un locus

Modèles déterministes

Soit une population de taille infinie composée d'individus diploides à reproduction sexuée. Cette population est supposée *isolée*, ce qui veut dire qu'elle ne reçoit aucun individu provenant d'une autre population. Elle se renouvelle entièrement à chaque génération t, $t \in \mathbb{N}$. On fait l'hypothèse de *panmixie*, selon laquelle tous les couples d'individus ont la même probabilité de s'unir pour donner naissance à un nouvel individu (la notion de sexe n'est pas prise en compte). On suppose enfin que la transmission des gènes est *mendélienne*, à savoir :

- un individu hétérozygote pour le locus étudié, c'est à dire de génotype i_1/i_2 avec $i_1 \neq i_2$, produit exactement 50% de gamètes i_1 et 50% de gamètes i_2.
- Pour deux parents donnés, toutes les combinaisons de gamètes ont la même probabilité d'être choisies pour donner naissance au nouvel individu.

Soit un locus biallélique (l'extension à un nombre quelconque d'allèles est immédiate). Notons $\Pi_{1/1}(t)$, $\Pi_{1/2}(t)$ et $\Pi_{2/2}(t)$ les fréquences des génotypes 1/1, 1/2 et 2/2 à la génération t. Ces quantités sont des variables déterministes à valeur dans $[0, 1]$. Pour tous les couples de génotypes parentaux possibles, le tableau 2.1 donne leurs fréquences et les probabilités des différents génotypes engendrés. On peut en déduire les fréquences des génotypes à la génération $t + 1$

$$\begin{aligned} \Pi_{1/1}(t+1) &= \Pi_{1/1}(t)^2 + \Pi_{1/1}(t)\Pi_{1/2}(t) + \frac{1}{4}\Pi_{1/2}(t)^2 \\ &= (\Pi_{1/1}(t) + \frac{1}{2}\Pi_{1/2}(t))^2 \end{aligned}$$

Et en suivant le même raisonnement

$$\begin{aligned} \Pi_{1/2}(t+1) &= (\Pi_{1/1}(t) + \frac{1}{2}\Pi_{1/2}(t))(\frac{1}{2}\Pi_{1/2}(t) + \Pi_{2/2}(t)) \\ \Pi_{2/2}(t+1) &= (\frac{1}{2}\Pi_{1/2}(t) + \Pi_{2/2}(t))^2 \end{aligned}$$

On peut en tirer deux conclusions. Premièrement, la fréquence des allèles dans la population est constante. Par exemple, si $\Pi_1(t)$ est la fréquence de l'allèle 1 à la génération

t, on a

$$\begin{aligned}\Pi_1(t+1) &= \Pi_{1/1}(t+1) + \frac{1}{2}\Pi_{1/2}(t+1) \\ &= (\Pi_{1/1}(t) + \frac{1}{2}\Pi_{1/2}(t))^2 + \frac{1}{2}(\Pi_{1/1}(t) + \frac{1}{2}\Pi_{1/2}(t))(\frac{1}{2}\Pi_{1/2}(t) + \Pi_{2/2}(t)) \\ &= \Pi_{1/1}(t) + \frac{1}{2}\Pi_{1/2}(t) \\ &= \Pi_1(t)\end{aligned}$$

Deuxièmement, les fréquences des génotypes sont simplement le produit des fréquences des allèles correspondants, puisque en regroupant les deux résultats précédents on trouve

$$\Pi_{1/1}(t+1) = (\Pi_{1/1}(t) + \frac{1}{2}\Pi_{1/2}(t))^2 = \Pi_1^2(t+1)$$

Cette propriété porte le nom d'*équilibre d'Hardy-Weinberg*. Elle est importante car elle nous permet de ne pas tenir compte du caractère diploïde de la population et de la modéliser comme un ensemble d'allèles. C'est ce que nous ferons dans la suite de ce chapitre. De même, pour des modèles à plusieurs loci, la population sera modélisée comme un ensemble d'haplotypes.

Génotypes des parents	Fréquence	Génotypes formés		
		1/1	1/2	2/2
$1/1 \times 1/1$	$\Pi_{1/1}(t)^2$	1	0	0
$1/1 \times 1/2$	$2\Pi_{1/1}(t)\Pi_{1/2}(t)$	0.5	0.5	0
$1/1 \times 2/2$	$2\Pi_{1/1}(t)\Pi_{2/2}(t)$	0	1	0
$1/2 \times 1/2$	$\Pi_{1/2}(t)^2$	0.25	0.5	0.25
$1/2 \times 2/2$	$2\Pi_{1/2}(t)\Pi_{2/2}(t)$	0	0.5	0.5
$2/2 \times 2/2$	$\Pi_{2/2}(t)^2$	0	0	1

TAB. 2.1 – Probabilité d'apparition des différents génotypes en fonction des génotypes parentaux

On peut compléter ce modèle pour tenir compte des mutations entre allèles qui se produisent parfois au cours de la méiose. Par exemple, imaginons qu'une fraction μ_1 d'allèles 2 se transforme en allèles 1 à chaque génération, et qu'une fraction μ_2 d'allèles 1 se transforme de même en allèles 2. La fréquence de l'allèle 1 évolue alors d'une génération à l'autre selon l'équation

$$\Pi_1(t+1) = (1-\mu_2)\Pi_1(t) + \mu_1\Pi_2(t)$$

On obtient alors facilement une expression de $\Pi_1(t)$ pour tout t en fonction de $\Pi_1(0)$ et des paramètres de mutation. Si $\mu_1 + \mu_2 < 1$ (ce qui est toujours le cas en pratique car les taux de mutations sont des quantités très faibles, au plus d'ordre 10^{-3}) les fréquences tendent vers une valeur fixe quand t tend vers l'infini. On a en effet

$$\lim_{t\to\infty} \Pi_1(t) = \Pi_1(\infty) = \frac{\mu_1}{\mu_1 + \mu_2}$$

On peut enfin inclure des effets de sélection qui interviennent entre la création de l'individu et sa maturité. Ici on considère à nouveau des individus diploïdes, en supposant l'équilibre d'Hardy-Weinberg. On introduit les coefficients de *viabilités* $\omega_{1/1}$, $\omega_{1/2}$ et $\omega_{2/2}$ des génotypes 1/1, 1/2 et 2/2. Ces coefficient permettent de quantifier la capacité de survie d'un individu en fonction de son génotype. On obtient alors la relation

$$\Pi_1(t+1) = \frac{\Pi_1^2(t)\omega_{1/1} + \Pi_1(t)\Pi_2(t)\omega_{1/2}}{\overline{\omega}}$$

où $\overline{\omega} = \Pi_1^2(t)\omega_{1/1} + 2\Pi_1(t)\Pi_2(t)\omega_{1/2} + \Pi_2^2(t)\omega_{2/2}$ est la viabilité moyenne dans la population. Suivant les valeurs des coefficients de viabilité, on peut aboutir à la *fixation* d'un des deux allèles (c'est à dire à la disparition de l'autre) ou à un état d'équilibre stable entre les deux fréquences. C'est par exemple le cas si la sélection avantage les individus hétérozygotes, c'est à dire si $0 < \omega_{1/1}, \omega_{2/2} < \omega_{1/2}$. On parle alors de *sélection balancée*, et l'état d'équilibre est donné par

$$\lim_{t\to\infty} \Pi_1(t) = \Pi_1(\infty) = \frac{\omega_{2/2} - \omega_{1/2}}{\omega_{1/1} + \omega_{2/2} - 2\omega_{1/2}}$$

Les effets de mutations et de sélection peuvent évidemment être cumulés. Pour simplifier l'exposé, la présentation des modèles qui suivent se fera dans le cas où il n'y a ni sélection ni mutation.

Modèle de Wright-Fisher

Le modèle déterministe décrit ci-dessus suppose une population de taille infinie. Il ne tient pas compte de la *dérive génétique*, c'est à dire du fait que dans une population de taille finie, même sans mutation et sans sélection, les fréquences alléliques puissent tout de même évoluer par le seul fait du hasard. Le *modèle de Wright-Fisher* (Wright, 1931), introduit par ces deux auteurs dans les années 1930, remédie à ce problème.

Les hypothèses générales sont les mêmes que précédemment mais la population est cette fois de taille finie et constante $2N$ ($2N$ est le nombre d'allèles). Les allèles formant la génération $t+1$ sont issus d'un tirage aléatoire avec remise parmi les allèles de la

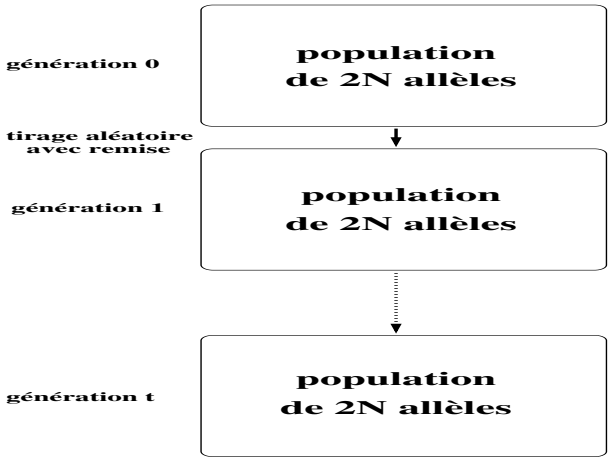


FIG. 2.1 – Évolution d'une population sous le modèle de Wright-Fisher

génération t (cf figure 2.1). Formellement, posons $X(t)$ le nombre d'allèles 1 présents à la génération t. On retrouve la fréquence par $\Pi(t) = X(t)/2N$. Sous le modèle de Wright-Fisher, le processus $\{X(t)\}_{t\in\mathbb{N}}$ est une chaîne de Markov à valeurs dans $\{0, 1, \ldots, 2N\}$. La loi de $X(t+1)$ sachant $X(t)$ est une loi binomiale $\mathcal{B}(2N, \Pi(t))$. On a par exemple

$$\mathbb{E}[X(t+1) \mid X(t)] = 2N\Pi(t) = X(t)$$

et

$$\mathrm{Var}[X(t+1) \mid X(t)] = 2N\Pi(t)(1 - \Pi(t))$$

En utilisant les propriétés des probabilités conditionnelles, on obtient des relations de récurrence qui conduisent finalement à $\forall t \in \mathbb{N}$

$$\mathbb{E}[X(t)] = \mathbb{E}[X(0)]$$

et

$$\mathrm{Var}(X(t)) = \mathbb{E}[X(0)](2N - \mathbb{E}[X(0)])(1 - (1 - 1/2N)^t) + (1 - 1/2N)^t \mathrm{Var}(X(0))$$

La loi de $X(t)$ est d'ailleurs entièrement calculable en utilisant l'expression de la matrice de transition. Mais pour des grandes populations, la taille de l'espace d'états rend ces calculs peu commodes. Prendre la limite quand N tend vers l'infini redonne le modèle déterministe de la section précédente. Pour simplifier les calculs sans perdre l'effet de dérive génétique, on utilise généralement une approximation par un processus de diffusion.

Modèle de diffusion

L'approximation du processus de Wright-Fisher par un processus de diffusion a été proposée par Kimura (1964). Elle correspond à un changement d'échelle de temps. Définissons en effet, pour tout $\tau \geq 0$, $Y^N(\tau) = \Pi([2N\tau]) = X([2N\tau])/2N$, où [ ] désigne la partie entière d'un réel. On montre -se reporter au chapitre 5 pour plus de détails- que quand N tend vers l'infini le processus $\{Y^N(\tau)\}_{\tau \geq 0}$ converge en distribution vers un processus de diffusion $\{Y(\tau)\}_{\tau \geq 0}$ à valeurs dans $[0,1]$. Ce processus est caractérisé par les coefficients

$$b(y) = \lim_{h \to 0} \frac{1}{h} \mathbb{E}[Y(\tau + h) - Y(\tau) \mid Y(\tau) = y] = 0$$

et

$$a^2(y) = \lim_{h \to 0} \frac{1}{h} \mathbb{E}[(Y(\tau + h) - Y(\tau))^2 \mid Y(\tau) = y] = y(1-y)$$

$b(.)$ est la *dérive infinitésimale* du processus et $a^2(.)$ sa *variance infinitésimale.*

Le processus limite $Y(.)$ est utilisé pour approcher le processus de Wright-Fisher pour N grand. Cela permet notamment d'obtenir des expressions explicites de l'espérance du temps de fixation d'un allèle ou de la densité de transition de la fréquence y^0 à la fréquence y (Kimura, 1955)

$$f(y^0, y, \tau) = y^0(1-y^0) \sum_{n=0}^{+\infty} C_n e^{-\frac{1}{2}(n+1)(n+2)\tau} P_n^{1,1}(1-2y^0) P_n^{1,1}(1-2y),\ 0 < y^0, y < 1,\ \tau > 0$$

où $C_n = (2n+3)(n+2)/(n+1)$ et où les $\{P_n^{1,1}\}_{n \in \mathbb{N}}$ sont les polynômes de Jacobi, qui forment une famille orthogonale sur $[-1,1]$ par rapport à la mesure $(1-y)(1+y)$. De nombreux autres exemples d'applications sont proposés dans (Karlin et Taylor, 1981).

2.1.2 Modèles à plusieurs loci

Pour la cartographie génétique, l'évolution des fréquences alléliques à un locus isolé ne suffit pas. Il faut s'intéresser à l'évolution conjointe des fréquences au niveau de plusieurs loci. On étudie donc les fréquences des haplotypes, et on doit maintenant tenir compte des recombinaisons entre loci.

Modèle de Wright-Fisher : cas général

Soient L loci ($L \geq 2$) d'une carte génétique, indicés en fonction de leur ordre sur cette carte (le premier locus rencontré a l'indice 1, le second à l'indice 2 etc). Chaque locus l a I_l allèles. On peut donc former $J = \prod_{l=1}^{L} I_l$ haplotypes à partir de ces loci, et les effectifs

de tous ces haplotypes sont contenus dans le vecteur $X(t)$. Comme précédemment, on introduit aussi le vecteur des fréquences $\Pi(t) = X(t)/2N$. Pour $l = 2, \ldots, L$, on note c_{l-1} le taux de recombinaison entre les loci $l-1$ et l.

On a toujours un modèle de chaîne de Markov où $X(t+1)$ suit, conditionnellement à $X(t)$, une loi multinomiale $(2N, p)$. p est un vecteur de taille J qui dépend de $X(t)$ et des taux de recombinaison entre loci. Pour L quelconque, il n'y a pas d'expression générale simple de p. Mais on peut décrire la loi de $X(t+1)$ sachant $X(t)$ par le schéma de simulation suivant :

1. Initialisation : $X(t+1)$ est le vecteur de taille J identiquement nul.
2. Pour n allant de 1 à $2N$:
   (a) Tirer deux haplotypes h et h' selon une loi multinomiale $(2, \Pi(t))$. On note $h = (i_1, \ldots, i_L)$ et $h' = (i'_1, \ldots, i'_L)$.
   (b) Choisir l'allèle i_1 avec une probabilité $1/2$ et l'allèle i'_1 sinon.
   (c) Pour l allant de 2 à L :
   - Si on vient de tirer i_{l-1}, choisir l'allèle i_l avec la probabilité $1 - c_{l-1}$ et l'allèle i'_l sinon.
   - Si on vient de tirer i'_{l-1}, choisir l'allèle i_l avec la probabilité c_{l-1} et l'allèle i'_l sinon.
3. Les L allèles tirés constituent l'haplotype d'un nouvel individu. Incrémenter de 1 la coordonnée de X(t+1) correspondante.

Ce procédé de simulation, itéré génération après génération, permet d'étudier le processus de Wright-Fisher dans le cadre le plus général.

Exemple pour deux loci

Le cas particulier du modèle à deux loci, défini par Karlin et McGregor (1968), a été souvent étudié. Dans ce cas, la probabilité p_{i_1,i_2} d'obtenir un haplotype (i_1, i_2) à la génération $t+1$ en utilisant l'algorithme ci-dessus vaut (cf annexe B.1)

$$p_{i_1,i_2} = (1-c)\Pi_{i_1,i_2}(t) + c\Pi_{i_1,.}(t)\Pi_{.,i_2}(t) \qquad i_1 = 1, \ldots, I_1,\ i_2 = 1, \ldots, I_2 \tag{2.1}$$

où $\Pi_{i_1,.}(t) = \sum_{i_2=1}^{I_2} \Pi_{i_1,i_2}(t)$ et $\Pi_{.,i_2}(t) = \sum_{i_1=1}^{I_1} \Pi_{i_1,i_2}(t)$ sont les fréquences marginales respectives des allèles i_1 au locus 1 et i_2 au locus 2 à la génération t, et où c est le taux de recombinaison entre les loci.

Si $c = 0$, les deux loci sont totalement liés et le modèle s'étudie comme un modèle à un locus. Si $c = 1/2$ les loci sont indépendants et on se ramène donc facilement à

l'étude des fréquences alléliques à chaque locus. Dans les autres cas ($0 < c < 1/2$), l'étude de $\{\Pi(t)\}_{t\in\mathbb{N}}$ n'est pas très facile. Ceci est dû à la forme de la probabilité de transition p_{i_1,i_2} qui, contrairement à ce qui se passait pour le modèle à un locus, est non linéaire et fait intervenir d'autres termes en plus de $\Pi_{i_1,i_2}(t)$. Aussi les résultats concernant ce modèle concernent principalement des moments particuliers du processus, qui quantifient la corrélation entre les deux loci. Nous y reviendrons dans le chapitre 3.

Comme dans le cas du modèle à un locus, il est possible d'approcher le processus de Wright-Fisher par un processus de diffusion. Cette diffusion a été obtenue par Ethier et Nagylaki (1989), et sera présentée au début du chapitre 5. L'objet de ce chapitre sera de calculer, dans le cadre de cette diffusion, la loi des fréquences d'haplotypes pour un modèle à deux loci.

Remarque 3 *Nous utiliserons au chapitre 8 un cas particulier de modèle de Wright-Fisher à trois loci, dans lequel le locus central est un QTL biallélique et les deux loci extrèmes sont des marqueurs multialléliques. Nous verrons que dans ce cas on peut également obtenir une expression simple du vecteur de probabilités p.*

2.2 Arbre de coalescence pour un échantillon

Pour l'analyse statistique d'un échantillon d'individus, il n'est pas forcément nécessaire de retracer l'histoire génétique de toute la population dont est issu cet échantillon. En effet les seuls événements génétiques ayant un impact sur notre observation sont ceux qui se sont produits parmi les ancêtres des individus de l'échantillon. Cette idée est à l'origine du développement de nombreux modèles permettant de décrire *l'ascendance*, autrement dit l'arbre des ancêtres, d'un échantillon d'haplotypes. Quelques uns de ces modèles, basés sur la notion de *coalescence*, sont présentés ici.

2.2.1 Modèle sans recombinaison

Soient N_s haplotypes issus d'une population de taille $2N$ constante au cours du temps et suivant le modèle de Wright-Fisher sans recombinaison présenté en 2.1.1. Rappelons qu'il s'agit d'un modèle d'haplotypes ; par conséquent dans toute cette section, les termes "parent", "enfant", ou "ancêtre" désignent des haplotypes et non pas des individus diploïdes. Nous considérons cette fois les générations dans le sens inverse, c'est-à-dire en direction du passé : $t = 0$ correspond à la génération actuelle, $t = 1$ à la génération des parents, et ainsi de suite. Puisque à une génération donnée chaque haplotype est l'image exacte d'un des haplotypes de la génération précédente, il est clair que le nombre d'ancêtres

de nos N_s haplotypes à $t = 0$ va diminuer de façon aléatoire au cours du temps jusqu'à ce qu'on ait plus qu'un seul ancêtre commun à l'échantillon. Il y a diminution stricte du nombre d'ancêtres quand plusieurs ancêtres sont issus d'un même parent à la génération précédente. On parle alors d'*événement de coalescence* entre les différentes lignées ancestrales. Les instants de coalescence et le choix des lignées qui coalescent constituent l'arbre de coalescence de l'échantillon. La loi de ces arbres est généralement représentée par le modèle de Kingman (1982a, 1982b), qui est présenté ici.

Sous le modèle de Wright-Fisher, le nombre d'enfants issus d'un parent donné suit une loi binomiale de paramètre $(2N, 1/2N)$, car $1/2N$ est la fréquence de ce parent. Vu dans l'autre sens, ceci revient à ce chaque enfant choisisse selon une loi uniforme un des $2N$ parents. Par conséquent, la probabilité $q(n)$ que les n haplotypes $(n = 1, \ldots, N_s)$ constituant les ancêtres de l'échantillon à la génération t n'aient pas d'ancêtre commun à la génération précédente vaut :

$$\begin{aligned} \frac{2N(2N-1)\cdots(2N-n+1)}{2N^n} &= \prod_{i=1}^{n-1}(1-\frac{i}{2N}) \\ &= 1-\frac{n(n-1)}{4N}+O(\frac{1}{N^2}) \end{aligned}$$

Et le nombre de générations $G(n, N)$ nécessaires à l'obtention d'un événement de coalescence suit une loi géométrique de paramètre $1 - q(n)$. Soit :

$$\begin{aligned} \mathbb{P}(G(n,N)=t) &= (q(n))^{t-1}(1-q(n)) \\ &= (1-\frac{n(n-1)}{4N}+O(\frac{1}{N^2}))^{t-1}(\frac{n(n-1)}{4N}+O(\frac{1}{N^2})) \end{aligned}$$

Le modèle de Kingman, comme le modèle de diffusion, est un modèle limite obtenu en considérant le changement de temps $\tau = t/2N$ et en faisant tendre N vers l'infini. Il est important de noter que pour ce processus limite la coalescence de 3 lignées ou plus est de probabilité nulle. Un événement de coalescence correspond donc à la fusion d'exactement deux lignées, et les instants $G(n)$ de ces événements suivent des lois exponentielles de moyenne $\frac{2}{n(n-1)}$. La densité de $G(n)$ vaut donc

$$f_n(\tau) = \frac{n(n-1)}{2} e^{\frac{n(n-1)}{2}\tau}$$

Ce modèle fournit un moyen simple et très rapide de simuler l'ascendance d'un échantillon de N_s haplotypes issus d'une population très grande.
Pour n allant de N_s à 2 :

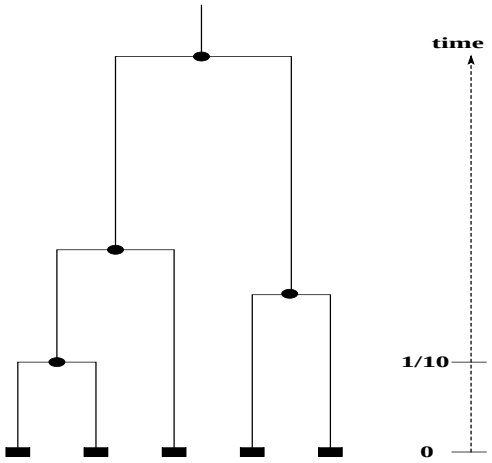


FIG. 2.2 – Arbre de coalescence pour un échantillon de taille $N_s = 5$ haplotypes. Les branches de l'arbre représentent les lignées ancestrales, et les cercles noirs les événements de coalescence entre ces lignées. Les carrés noirs désignent les haplotypes de l'échantillon. L'échelle de temps à droite indique que l'événement de coalescence le plus récent s'est produit au temps $\tau = 1/10$, c'est à dire il y a $2N/10$ générations.

1. Simuler une loi exponentielle de moyenne $\frac{2}{n(n-1)}$.
2. Choisir uniformément celui des $\frac{n(n-1)}{2}$ couples d'haplotypes qui coalesce.

Un exemple de réalisation du coalescent est illustré par la figure 2.2.

Les propriétés et extensions du coalescent ont été largement étudiées depuis les premiers articles de Kingman. On pourra se reporter à (Nordborg, 2001) ou (Tavaré et Zeitouni, 2004) pour plus de détails à ce sujet.

2.2.2 Modèle avec recombinaison

Une extension importante du modèle de coalescent concerne la prise en compte de la recombinaison dans l'histoire des N_s haplotypes observés. Comme nous l'avons vu avec l'algorithme de simulation de la section 2.1.2, ceci a pour conséquence qu'un haplotype pris à une génération donnée n'a plus un mais a priori deux parents à la génération précédente. La généalogie d'un échantillon d'individus ne peut donc plus être représentée par un arbre mais par un graphe. Deux lignées peuvent fusionner en une seule quand il y a coalescence, mais une lignée peut également se séparer en deux quand il y a recombinaison. A l'image de ce qui a été présenté dans le cas sans recombinaison, la limite du processus de Wright-Fisher présenté en 2.1.2 pour $\tau = 1/2N$ et $N \to \infty$ est l'*ancestral recombination graph* de Griffiths et Marjoran (1996). Il fait intervenir le taux de recombinaison limite $\rho = lim_{N\to\infty} 2Nc$, c étant le taux de recombinaison entre les deux loci extrèmes des

haplotypes.

Sous ce modèle, le taux d'un événement de recombinaison est de ρ pour chaque lignée, et on peut donc simuler la généalogie des N_s haplotypes de la manière suivante.

$n = N_s$. Tant que $n > 1$:

1. Simuler U, une loi exponentielle de moyenne $\frac{2}{n(n-1)}$.
2. Simuler V, une loi exponentielle de moyenne $n\rho$.
3. – Si $U \leq V$, choisir au hasard celui des $\frac{n(n-1)}{2}$ couples d'haplotypes qui coalesce.
   – Si $U > V$:
   (a) Choisir uniformément celle des n lignées qui se sépare.
   (b) Choisir uniformément sur l'haplotype le point de crossing-over.

La probabilité qu'il y ait deux crossing-over au cours d'un même événement de recombinaison est nulle.

Wiuf et Hein (1997) ont montré que l'état "un seul ancêtre" était toujours atteint en dépit du fait que le nombre de lignées puisse parfois augmenter. Mais le temps nécessaire pour atteindre cet état est évidemment plus long que pour le modèle sans recombinaison, et ce d'autant plus que ρ est grand. Une fois le graphe construit, on peut en déduire l'arbre de coalescence pour chaque locus. Il suffit pour cela de partir des feuilles de l'arbre (les N_s observations) et de suivre les branches en choisissant, quand il y a recombinaison, celle des deux branches qui contient le locus étudié (cf figure 2.3). Cette distinction se fait grâce à l'information ajoutée dans l'arbre lors de l'étape 3 (b) de la simulation. Les lois des arbres obtenus pour différents loci sont naturellement corrélées.

Références

Ethier, S., et Nagylaki, T. (1989). Diffusion approximations of the two-locus wright-fisher model. *J. Math. Biol.*, *27*, 17-28.

Griffiths, R., et Marjoran, P. (1996). Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology*, *3*, 479-502.

Karlin, S., et McGregor, J. (1968). Rates and probabilities of fixation for two locus random mating finite populations without selection. *Genetics*, *58*, 141-159.

Karlin, S., et Taylor, H. (1981). *A second course in stochastic processes*. Academic Press, Inc.

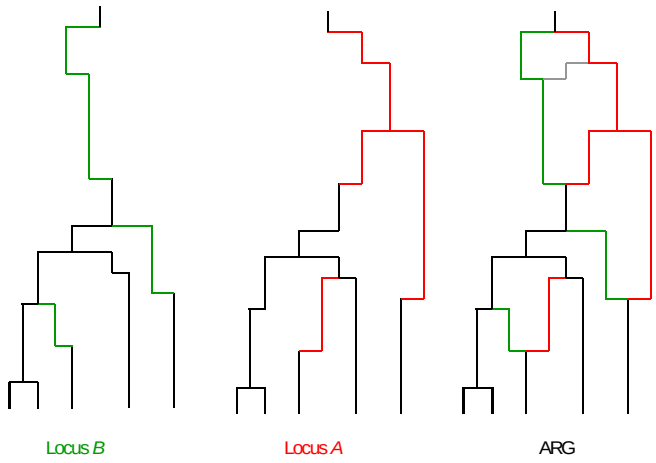


FIG. 2.3 – Réalisation d'un ancestral recombination graph pour un échantillon de taille $N_s = 5$ chromosomes et arbres de coalescents marginaux pour deux loci A et B

Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Nat. Acad. Sci. USA*, *41*, 144-150.

Kimura, M. (1964). Diffusion models in population genetics. *J. Appl. Probab.*, *1*, 177-232.

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

Kingman, J. (1982b). On the genealogy of large populations. *Applied Probability Trust*(19A), 27-43.

Nordborg, M. (2001). Coalescent theory. In D. Balding, M. Bishop, et C. Cannings (Eds.), *Handbook of statistical genetics* (p. 179-212). Wiley.

Tavaré, S., et Zeitouni, O. (2004). *Lectures on probability theory and statistics. Ecole d'été de Probabilités se Saint-Flour XXXI-2001* (Vol. 1837). Springer Verlag, New York.

Wiuf, C., et Hein, J. (1997). On the number of ancestors to a dna sequence. *Genetics*, *147*, 1459-1468.

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

Chapitre 3

Déséquilibre de liaison

La notion de déséquilibre de liaison entre deux allèles situés à des loci distincts fait référence à la corrélation entre ces allèles. Autrement dit il s'agit de savoir si ces allèles ont plutôt tendance à être hérités ensemble ou pas. L'objectif de ce chapitre est de montrer pourquoi cette notion est utilisée en cartographie génétique. Nous commencerons par définir précisément ce qu'est le déséquilibre de liaison et comment le mesurer. Puis nous illustrerons par un exemple simple le lien entre taux de recombinaison et déséquilibre de liaison, ce qui permettra d'introduire l'idée de cartographie génétique par déséquilibre de liaison. Dans une dernière partie nous nous intéresserons à une question classique : la présence de déséquilibre de liaison implique-t-elle que les fréquences d'haplotypes soient en régime transitoire ?

3.1 Définition et mesures

L'expression *déséquilibre de liaison* traduit, comme on peut le penser, un écart par rapport à une certaine situation d'équilibre, qui veut que pour deux allèles i_1 et i_2 situés à des loci différents, on ait :

$$\Pi_{i_1,i_2}(t) = \Pi_{i_1,.}(t)\Pi_{.,i_2}(t)$$

où $\Pi_{i_1,i_2}(t)$, $\Pi_{i_1,.}(t)$ et $\Pi_{.,i_2}(t)$ désignent respectivement les fréquences de l'haplotype (i_1, i_2), de l'allèle i_1 du premier locus et de l'allèle i_2 du second locus dans la population à la génération t. La manière la plus simple de quantifier le déséquilibre de liaison est donc la mesure

$$D_{i_1,i_2}(t) = \Pi_{i_1,i_2}(t) - \Pi_{i_1,.}(t)\Pi_{.,i_2}(t) \tag{3.1}$$

On parle d'*association positive* entre i_1 et i_2 si $D_{i_1,i_2}(t) > 0$, ce qui traduit le fait que les deux allèles ont tendance à être hérités ensemble ; l'association est au contraire *négative* si $D_{i_1,i_2}(t) < 0$.

Il existe aussi d'autres mesures du déséquilibre de liaison. Elles dérivent généralement de $D_{i_1,i_2}(t)$ mais y ajoutent un terme de normalisation. On peut citer notamment la mesure

$$r^2_{i_1,i_2}(t) = \frac{D^2_{i_1,i_2}(t)}{\Pi_{i_1,.}(t)(1-\Pi_{i_1,.}(t))\Pi_{.,i_2}(t)(1-\Pi_{.,i_2}(t))}$$

qui correspond à la statistique de test du χ^2 pour des tables de contingence 2×2.

Pour des loci bialléliques, on montre facilement que

$$D_{1,1}(t) = -D_{1,2}(t) = -D_{2,1}(t) = D_{2,2}(t)$$

Dans ce cas on peut donc parler du déséquilibre de liaison entre les loci sans préciser les allèles concernés, et on note simplement $|D|$ ou r^2 les mesures correspondantes. Pour les loci multialléliques, il existe des mesures permettant de définir le déséquilibre de liaison entre loci, comme par exemple

$$D^2(t) = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} D_{i_1,i_2}(t)$$

où I_1 est le nombre d'allèles au locus 1 et I_2 le nombre d'allèles au locus 2, ou

$$\Delta^2(t) = \frac{D^2(t)}{(1-\sum_{i_1=1}^{I_1} \Pi_{i_1,.})(1-\sum_{i_2=1}^{I_2} \Pi_{.,i_2})}$$

Signalons toutefois que dans la suite de ce manuscrit, nous parlerons généralement du déséquilibre de liaison entre deux loci pour désigner en fait les déséquilibres de liaison entre les allèles des deux loci, sans que cela fasse référence à une mesure particulière.

Une revue détaillée des mesures de déséquilibre de liaison et de leurs propriétés est fournie par Cierco-Ayrolles et al. (2004). Ce travail, auquel j'ai contribué, est inclus en annexe.

3.2 Influence du taux de recombinaison sur le déséquilibre de liaison

Plaçons-nous dans le cadre du modèle de Wright-Fisher à deux loci décrit en 2.1.2. Le taux de recombinaison entre les loci est c et la population est de taille $2N$. La proposition ci-dessous donne une idée de l'évolution du déséquilibre de liaison au cours des générations.

Proposition 1 *Soit $D_{i_1,i_2}(t)$ la mesure de déséquilibre de liaison définie par l'équation (3.1), et soit $T \in \mathbb{N}$. Dans le modèle de Wright-Fisher à deux loci défini en 2.1.2 on a :*

- *(i)* $\mathbb{E}[D_{i_1,i_2}(t+1)] = (1-c)(1-\frac{1}{2N})\mathbb{E}[D_{i_1,i_2}(t)],\ t = 1, \ldots, T-1$
- *(ii)* $\mathbb{E}[D_{i_1,i_2}(T)] = \mathbb{E}[D_{i_1,i_2}(0)](1-c)^T(1-\frac{1}{2N})^T$

Le (ii) découle directement du (i) par récurrence. Pour prouver le (i), on commence par établir deux résultats simples sur l'espérance et la variance des fréquences d'haplotypes en utilisant les propriétés des lois multinomiales. On trouve que

$$\begin{aligned}\mathbb{E}[\Pi_{i_1,i_2}(t+1) \mid X(t)] &= \frac{1}{2N}\mathbb{E}[X_{i_1,i_2}(t+1) \mid X(t)] \\ &= p_{i_1,i_2} \\ &= (1-c)\Pi_{i_1,i_2}(t) + c\Pi_{i_1,.}(t)\Pi_{.,i_2}(t)\end{aligned}$$

Pour $\mathrm{Cov}(\Pi_{i_1,.}(t+1), \Pi_{.,i_2}(t+1) \mid X(t))$, que l'on note ici $\rho(t)$ pour alléger les formules, on obtient

$$\begin{aligned}
\rho(t) &= \frac{1}{(2N)^2}\mathrm{Cov}(X_{i_1,.}(t+1), X_{.,i_2}(t+1) \mid X(t)) \\
&= \frac{1}{(2N)^2}\mathrm{Cov}\left(\sum_{j_2=1}^{I_2} X_{i_l,j_2}(t+1), \sum_{j_1=1}^{I_1} X_{j_1,i_2}(t+1) \mid X(t)\right) \\
&= \frac{1}{2N}\left(p_{i_1,i_2}(1-p_{i_1,i_2}) - \sum_{(j_1,j_2)\neq(i_1,i_2)} p_{i_1,j_2}p_{j_1,i_2}\right) \\
&= \frac{1}{2N}\left(p_{i_1,i_2} - \sum_{(j_1,j_2)} p_{i_1,j_2}p_{j_1,i_2}\right) \\
&= \frac{1}{2N}\left(p_{i_1,i_2} - (\sum_{j_1} p_{j_1,i_2})(\sum_{j_2} p_{i_1,j_2})\right) \\
&= \frac{1}{2N}\left(p_{i_1,i_2} - \Pi_{i_1,.}(t)\Pi_{.,i_2}(t)\right) \\
&= \frac{1}{2N}(1-c)D_{i_1,i_2}(t)
\end{aligned}$$

On en déduit alors une expression de $d(t) = \mathbb{E}[D_{i_1,i_2}(t+1) \mid X(t)]]$. En effet

$$\begin{aligned}
d(t) &= \mathbb{E}[\Pi_{i_1,i_2}(t+1) \mid X(t)] - \mathbb{E}[\Pi_{i_1,.}(t+1)\Pi_{.,i_2}(t+1) \mid X(t)] \\
&= (1-c)\Pi_{i_1,i_2}(t) + c\Pi_{i_1,.}(t)\Pi_{.,i_2}(t) - \mathbb{E}[\Pi_{i_1,.}(t+1)\Pi_{.,i_2}(t+1) \mid X(t)] \\
&= (1-c)D_{i_1,i_2}(t) + (\Pi_{i_1,.}(t)\Pi_{.,i_2}(t) - \mathbb{E}[\Pi_{i_1,.}(t+1)\Pi_{.,i_2}(t+1) \mid X(t)]) \\
&= (1-c)D_{i_1,i_2}(t) + \mathbb{E}[\Pi_{i_1,.}(t+1) \mid X(t)]\mathbb{E}[\Pi_{.,i_2}(t+1) \mid X(t)] \\
&- \mathbb{E}[\Pi_{i_1,.}(t+1)\Pi_{.,i_2}(t+1) \mid X(t)] \\
&= (1-c)D_{i_1,i_2}(t) - \mathrm{Cov}(\Pi_{i_1,.}(t+1), \Pi_{.,i_2}(t+1) \mid X(t)) \\
&= (1-c)(1-\frac{1}{2N})D_{i_1,i_2}(t)
\end{aligned}$$

L'espérance de cette formule redonne bien la relation (i).

La proposition 1 montre que pour une population isolée de taille $2N$ sans mutation ni sélection, le déséquilibre de liaison entre deux loci séparés par un taux de recombinaison c décroît en moyenne géométriquement à la vitesse $(1-c)(1-\frac{1}{2N})$. A moins que c soit très faible, on s'attend donc à observer peu de déséquilibre de liaison. La présence de déséquilibre de liaison entre deux loci peut par conséquent indiquer que ceux-ci sont

proches, et cette information est donc utile pour la cartographie. Nous présenterons au chapitre 4 plusieurs méthodes de cartographie permettant d'exploiter cette information.

3.3 Déséquilibre et loi stationnaire

Pour étudier une chaîne de Markov telle que celle décrite par le modèle de Wright-Fisher, on cherche généralement à caractériser sa loi stationnaire. Cela a t-il un sens dans le cas du déséquilibre de liaison qui, intuitivement et d'après la proposition du paragraphe précédent, est un phénomène typiquement transitoire ? En se basant sur le cas plus simple du modèle à un locus, nous verrons tout d'abord que le modèle de Wright-Fisher n'a pas forcément une unique loi stationnaire. Partant de cette constatation, nous reviendrons ensuite sur quelques résultats limites concernant le déséquilibre de liaison et nous nous interrogerons sur le sens qu'on peut leur attribuer.

3.3.1 Loi stationnaire pour le modèle à un locus

Considérons le modèle de Wright-Fisher à un locus décrit en 2.1.1. La chaîne $\{\Pi(t)\}_{t\in\mathbb{N}}$ possède deux états absorbants $\{0\}$ et $\{1\}$ tels que $\mathbb{P}(\Pi(\infty)=0 \mid \Pi(0)=\pi^0)=1-\pi^0$ et $\mathbb{P}(\Pi(\infty)=1 \mid \Pi(0)=\pi^0)=\pi^0$. En effet on montre facilement (Tavaré et Zeitouni, 2004) que $\forall j=1,\ldots,2N$, on a $\mathbb{P}(X(\infty)=2N \mid X(0)=j)=j/2N$. La chaîne $\{\Pi(t)\}_{t\in\mathbb{N}}$ admet donc deux lois stationnaires, qui sont les mesures de Dirac en 0 et en 1. En revanche le processus $\{\Pi(t) \mid 0<\Pi(t)<1\}_{t\in\mathbb{N}}$ admet une loi stationnaire unique. L'approximation par le processus de diffusion présenté en 2.1.1 permet de montrer que cette loi stationnaire est uniforme sur $]0,1[$. En effet on a vu que pour ce processus la densité de transition s'écrit

$$f(y^0,y,\tau)=y^0(1-y^0)\sum_{n=0}^{+\infty}C_n e^{-\frac{1}{2}(n+1)(n+2)\tau}P_n^{1,1}(1-2y^0)P_n^{1,1}(1-2y)$$

Quand $\tau\to\infty$ cette densité est équivalente à $6y^0(1-y^0)e^{-\tau}$ car c'est le terme en $n=0$ qui l'emporte. Or $P_0^{1,1}$ est constant égal à 1 et $C_0=6$. On peut montrer de même que

$$\begin{aligned}\int_0^1 f(y^0,y,\tau)dy &= 2y^0(1-y^0)\sum_{n=0}^{+\infty}\frac{4n+3}{2n+1}e^{-(2n+1)(n+1)\tau}P_{2n}^{1,1}(1-2y^0)\\ &\backsim 6y^0(1-y^0)e^{-\tau}\end{aligned}$$

On en déduit donc que

$$\lim_{\tau\to\infty}\frac{f(y^0,y,\tau)}{\mathbb{P}(0<Y(\tau)<1)}=1$$

Ajouter des mutations au modèle de Wright-Fisher en modifiant les probabilités de transition implique que les états $\{0\}$ et $\{1\}$ ne sont plus absorbants puisque tout allèle qui disparaît peut être recréé par mutation. On a donc également dans ce cas une loi stationnaire unique, et la densité du modèle de diffusion correspondant est également bien connue (Wright, 1949).

3.3.2 Loi limite du déséquilibre de liaison

Ces mêmes nuances se retrouvent pour l'étude du déséquilibre de liaison entre deux loci. Si on considère le modèle de Wright-Fisher sans mutation décrit en 2.1.2, il a plusieurs états absorbants et admet donc plusieurs lois stationnaires. Par exemple pour deux loci biallélіques, le processus considéré est $\Pi(t) = (\Pi_{1,1}(t), \Pi_{1,2}(t), \Pi_{2,1}(t), \Pi_{2,2}(t))$ et il a quatre états absorbants : $\{(1,0,0,0)\}$, $\{(0,1,0,0)\}$, $\{(0,0,1,0)\}$ et $\{(0,0,0,1)\}$. La loi limite de ce modèle quand t tend vers l'infini n'apporte pas vraiment d'information pour l'étude du déséquilibre de liaison car en pratique on ne s'intéresse jamais au déséquilibre de liaison entre deux loci s'ils ne sont pas polymorphes. Par exemple la proposition 1 indique que

$$\lim_{t\to\infty} \mathbb{E}[D_{1,1}(t)] = 0$$

Une formule similaire à celle de la proposition 1 a également été proposée pour $\mathbb{E}[D_{1,1}^2(t)]$ (Littler, 1973). Elle implique que

$$\lim_{t\to\infty} \mathbb{E}[D_{1,1}^2(t)] = 0$$

Toutefois ce résultat reflète juste le fait que $D_{1,1} = 0$ dans chacun des 4 états absorbants.

Pour l'étude du déséquilibre, il est donc plus judicieux de considérer la loi du processus sachant qu'aux deux loci aucun allèle n'a fixé. Cette condition est plus forte que le simple rejet des 4 états absorbants. Par exemple pour deux loci biallélіques, les états de type $\{(y, 1-y, 0, 0), 0 < y < 1\}$ doivent être rejetés car ils correspondent à la fixation de l'allèle 1 au premier locus. Cette condition de non fixation est implicite dans les raisonnements de Sved (1971) quand il montre que pour un modèle à deux loci biallélіques

$$\lim_{t\to\infty} \mathbb{E}[r^2(t)] = \frac{1}{1 + 4Nc\frac{(1-c/2)}{(1-c)^2}} \approx \frac{1}{1+4Nc}$$

En effet, la première partie de sa démonstration consiste à montrer que pour une population donnée, la probabilité Q_t que deux loci soient IBD depuis t générations (c'est à dire qu'il n'y ait pas eu de recombinaison entre ces loci au cours des t dèrnières générations)

peut être estimée par la mesure r^2. Il est clair que ceci n'a pas de sens si il y a eu fixation à un locus, puisque dans ce cas la mesure r^2 n'est même pas définie.

Bien qu'on n'ait pas d'expression exacte de $\mathbb{E}[D_{1,1}(t)]$ sachant qu'il n'y a pas encore eu fixation, on peut penser que cette quantité tend vers 0 quand t tend vers l'infini car la vitesse de convergence de $\mathbb{E}[D_{1,1}(t)]$ vers 0 est beaucoup plus rapide que la vitesse de convergence des fréquences alléliques ($\Pi_{1,1}(t)$, $\Pi_{1,2}(t)$, $\Pi_{2,1}(t)$, et $\Pi_{2,2}(t)$) vers 0. On peut donc parler d'une loi stationnaire dans laquelle le déséquilibre de liaison, en moyenne nul, se mesure par sa variance. La formule de Sved donne une approximation de cette variance, et illustre le rôle de la taille de la population N et du taux de recombinaison c : le déséquilibre de liaison est plus important dans les populations de petite taille, et pour des loci proches.

Si le modèle inclut des mutations, il existe bien une loi stationnaire unique dans laquelle on peut calculer l'espérance de différentes mesures de déséquilibre. On trouve alors effectivement que $\mathbb{E}[D_{i_1,i_2}(t)] \to 0$ quand $t \to \infty$ (Griffiths, 1981), et on peut obtenir d'autres résultats asymptotiques concernant les mesures des déséquilibres de liaison. Griffiths (1981) a notamment donné la limite de $\mathbb{E}[D^2(t)]$ pour deux loci multialléliques sous différents modèles de mutations classiques. D'autres auteurs (Hill, 1975; McVean, 2002) ont proposé des approximations pour la limite de $\mathbb{E}[\Delta^2(t)]$. Tous ces résultats ont été reportés par Cierco-Ayrolles et al. (2004) et sont présentés dans l'annexe A.

Avec ou sans mutations, une meilleure caractérisation du modèle de Wright-Fisher à deux loci semble nécessaire. La loi stationnaire, quand elle est unique, n'est pas forcément adaptée pour décrire le comportement de fréquences d'haplotypes qui est souvent remis en cause par des événements démographiques ou génétiques non pris en compte dans le modèle. Dans le chapitre 5 de ce manuscrit, nous présenterons une méthode visant à mieux évaluer le comportement transitoire du modèle de Wright-Fisher.

Références

Cierco-Ayrolles, C., Abdallah, J., Boitard, S., Chikhi, L., Rochambeau, H. de, Tsitrone, A., et al. (2004). On linkage disequilibrium measures : Methods and applications. In (Vol. 1, p. 151-180). Research Signpost, India.

Griffiths, R. (1981). Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.*, *19*, 169-186.

Hill, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in a finite population. *Theor. Popul. Biol.*, *8*, 117-126.

Littler, R. (1973). Linkage disequilibrium in two-locus, finite, random mating models without selection or mutation. *Theor. Popul. Biol.*, *4*, 259-275.

McVean, G. A. (2002). A genealogical interpretation of linkage disequilibrium. *Genetics*, *162*, 987-991.

Sved, J. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.*, *2*, 125-141.

Tavaré, S., et Zeitouni, O. (2004). *Lectures on probability theory and statistics. Ecole d'été de Probabilités se Saint-Flour XXXI-2001* (Vol. 1837). Springer Verlag, New York.

Wright, S. (1949). In G. Jepson, G. Simpson, et E. Mayr (Eds.), *Genetics, paleontology and evolution* (p. 365-389). Princeton University Press, Princeton, N.J.

Chapitre 4

Méthodes de cartographie fine

Nous avons vu au cours du chapitre précédent que le déséquilibre de liaison pouvait être utile dans le cadre de la cartographie génétique. En fait, une mesure du déséquilibre de liaison n'est autre qu'une statistique permettant de résumer l'information que nous fournissent les observations au sujet de l'histoire de la population étudiée. De manière générale, on définit la cartographie par déséquilibre de liaison comme l'ensemble des méthodes utilisant l'échantillon d'individus observés comme un témoin de l'histoire de la population. Ce chapitre a pour but de présenter l'état actuel des recherches dans le domaine de la cartographie fine par déséquilibre de liaison.

Depuis quelques années, la cartographie fine par déséquilibre de liaison est l'objet de très nombreux travaux de recherche. Nous essaierons de faire ressortir les différents types d'approches utilisés. La section 4.1 introduit les hypothèses générales de la cartographie fine par déséquilibre de liaison, et les hypothèses particulières du cas auquel nous nous sommes intéressés. Différentes approches sont ensuite présentées. En section 4.2, il s'agit de modèles linéaires simples visant à expliquer les phénotypes des individus par leurs génotypes aux marqueurs ; dans la section 4.3, les méthodes s'appuient sur des modèles de génétique des population. En section 4.4, nous donnons finalement les raisons qui nous ont conduits à étudier plus particulièrement l'une de ces approches.

4.1 Hypothèses générales

Soit un échantillon de N_s individus non apparentés. On dispose de leurs phénotypes z_n, $n = 1, \ldots, N_s$, et de leurs génotypes m_n, $n = 1, \ldots, N_s$ pour L marqueurs d'une carte génétique. Nous noterons l'ensemble de ces données $\mathcal{D} = \{(z_n, m_n), n = 1, \ldots, N_s\}$. Dans la zone délimitée par les L marqueurs, on suppose qu'il existe un unique gène ayant une influence sur les phénotypes observés. Notre objectif est d'estimer la position x de ce gène

sur la carte génétique. Suivant le problème posé, plusieurs hypothèses sont alors possibles concernant :

- les phénotypes : les z_n peuvent être vus comme les réalisations de variables discrètes (caractère mendelien) ou continues (QTL).
- l'origine des individus : on peut supposer qu'ils sont issus d'une seule population homogène ou bien de plusieurs populations. Cette structuration en populations peut être connue ou inconnue.
- l'échantillonage : les individus ne sont pas forcément tirés au hasard. Pour la recherche de QTL, on sélectionne parfois les individus ayant des valeurs phénotypiques extrêmes. Pour les gènes liés à des maladies rares, les individus malades sont toujours sur-représentés par rapport à leur fréquence dans la population (sans quoi on n'en aurait qu'un nombre très faible). Ce type d'études porte le nom de *cas-contrôle*.
- les génotypes : si un individu est hétérozygote, par exemple 1/2, les procédés biologiques connus à ce jour ne permettent pas de déterminer lequel des chromosomes homologues porte l'allèle 1 et lequel porte l'allèle 2. Quand plusieurs marqueurs sont analysés, on n'est donc pas toujours capable de déterminer *la phase* des génotypes, c'est à dire la paire d'haplotypes correspondante. Par exemple si deux marqueurs ont des génotypes 1/2, les haplotypes correspondants peuvent être $(1,1)/(2,2)$ ou $(1,2)/(1,2)$. Pourtant, il est fréquent de supposer que la phase est connue, car il existe des algorithmes de plus en plus fiables pour reconstruire les haploypes à partir des seuls génotypes (Niu, 2004; Marchini et al., 2006).

Dans ce manuscrit, nous nous intéressons à des caractères quantitatifs. De plus, les travaux présentés dans les parties II et III supposent un échantillonage uniforme et une population homogène. Ils nécessitent également la connaissance des haplotypes. L'état de l'art présenté dans ce chapitre est donc spécifique à ce type d'hypothèses, même si certaines des méthodes qui y seront évoquées ont été développées pour des cas légèrement différents.

Le travail présenté en partie IV sort du cadre de ce chapitre, car les individus observés sont issus de plusieurs populations distinctes et s'organisent en *familles nucléaires* de type père-mère-enfant. D'autre part l'objectif n'est pas d'estimer la position d'un QTL à l'intérieur d'une zone délimitée mais de tester la présence d'un QTL pour un grand nombre de loci répartis sur le génome. Une bibliographie rapide des méthodes adaptées à ces hypothèses sera proposée au chapitre 10.

4.2 Méthodes d'association

Une *méthode d'association* est une méthode qui, pour un marqueur ou un groupe de marqueurs donnés, cherche à tester s'ils ont un effet sur l'expression du caractère étudié. Pour estimer la position d'un QTL, la première idée proposée a été de tester individuellement l'effet de chaque marqueur de la carte et de retenir celui ayant la statistique de test la plus élevée (Boerwinkle et al., 1986, 1987). Pour un marqueur ayant I allèles, on peut par exemple adopter le modèle d'ANOVA

$$Z_{k,j} = m_k + \epsilon_{k,j},\ k = 1, \ldots, K,\ j = 1, \ldots, N_k \quad (M)$$

où m_k est l'effet du génotype k, N_k le nombre d'individus ayant le génotype k, et $K = \frac{I(I-1)}{2}$ le nombre de génotypes possibles au marqueur. On teste alors dans (M) l'hypothèse H_0 : "$m_1 = \ldots = m_K$". On peut également supposer que les allèles ont un effet additif et on obtient ainsi le modéle de régression

$$Z_n = \sum_{i=1}^{I} a_i x_{n,i} + \epsilon_n,\ n = 1, \ldots, N_s \quad (M')$$

où a_i est l'effet de l'allèle i, $x_{n,i}$ le nombre d'allèles i observés chez l'individu n ($x_{n,i} = 0, 1, 2$) et ϵ_n un bruit gaussien centré. On teste alors dans (M') l'hypothèse H_0 : "$a_1 = \ldots = a_I$". On peut noter que (M') est un sous-modèle de (M).

Le lien entre la valeur du test et la position du QTL a été justifié par Nielsen et Weir (1999) de la manière suivante : supposons qu'il y ait un QTL biallélique quelque part sur la carte génétique, et considérons un marqueur quelconque. L' effet d'un allèle de ce marqueur s'exprime comme une fonction simple de l'effet du QTL et du déséquilibre de liaison entre cet allèle et les allèles du QTL. Si plusieurs marqueurs sont étudiés successivement, ceux dont les allèles ont un effet important sont donc également ceux pour lesquels le déséquilibre de liaison avec le QTL est le plus grand. Estimer la position du QTL par la position du marqueur ayant la statistique de test la plus importante revient donc indirectement à faire ce que Devlin et Risch (1995) appellent de la *cartographie simple par déséquilibre de liaison*. Développée pour des gènes de maladie pour lesquels l'observation du phénotype d'un individu permet facilement de déduire le génotype de celui-ci, cette approche consiste à estimer la position du gène par la position du marqueur ayant le plus fort déséquilibre de liaison avec le gène.

Si on dispose de plusieurs marqueurs très proches (dans le même gène par exemple), on peut naturellement étendre le modèle (M) en considérant tous les génotypes possibles

pour le groupe de marqueurs et en attribuant un effet à chaque génotype, ou étendre le modèle (M') en attribuant un effet à chaque haplotype. Une autre solution consiste à supposer que les effets des marqueurs s'additionnent, et à modéliser l'effet de chaque marqueur à l'aide du modèle (M). Fan et Xiong (2002) ont étudié les propriétés d'un tel modèle dans le cas de deux marqueurs bialléliques, et Jung et al. (2005) ont étendu leurs résultats à un nombre quelconque de marqueurs. Pour de nombreux marqueurs ou des marqueurs ayant beaucoup d'allèles, on est néanmoins confronté à un problème de puissance dû au nombre trop important de paramètres à estimer. Une solution consiste alors à regrouper certains haplotypes en utilisant des méthodes de clustering (Tzeng et al., 2006; Molitor et al., 2003; Waldron et al., 2006).

4.3 Utilisation de modèles de populations

La structure de l'échantillon d'haplotypes observés (et non observés si on inclut l'allèle du QTL) n'est pas uniforme. Elle découle de l'histoire génétique de la population considérée. Quand un nouvel allèle au QTL apparaît dans une population (par mutation, migration, ...), il est associé à un unique haplotype au niveau des marqueurs étudiés. Sous l'effet des recombinaisons, cet haplotype initial va ensuite se dégrader au cours des générations, mais restera en principe plus fréquent que les autres, surtout pour les marqueurs très proches du QTL (Hästbacka et al., 1992; Jorde, 1995).

Modéliser cette histoire de manière explicite dans les méthodes de localisation permet de réduire le nombre de paramètres à estimer, et donc d'améliorer la précision des estimations. Les modèles utilisés sont ceux qui ont été décrits au chapitre 2, et plusieurs types de méthodes peuvent être distingués. Dans toutes ces méthodes, le calcul de la vraisemblance $\mathcal{L}(x \mid \mathcal{D})$, où x est la position du QTL et $\mathcal{D}$ est l'ensemble des données disponibles, défini en 4.1, joue un rôle central.

4.3.1 Méthodes basées sur les fréquences d'haplotypes

Une première façon d'aborder le problème est d'utiliser les variables cachées que sont les fréquences d'haplotypes dans la population, pour les haplotypes comprenant à la fois le QTL et les marqueurs. En effet pour un haplotype aux marqueurs h_n, le rapport de fréquences $\frac{\Pi_{Q,h_n}}{\Pi_{h_n}}$ donne la probabilité d'avoir un allèle Q au QTL sachant qu'on a l'haplotype h_n aux marqueurs. Cette quantité est importante puisque c'est l'allèle au QTL qui détermine la loi du phénotype Z_n. Si $\Pi(t)$ est le vecteur de toutes les fréquences d'haplotypes dans la population actuelle, il est donc facile d'obtenir la vraisemblance complète

$\mathcal{L}(x \mid \mathcal{D}, \Pi(t))$. On calcule donc la vraisemblance incomplète en prenant l'espérance par rapport à la loi de $\Pi(t)$:

$$\mathcal{L}(x \mid \mathcal{D}) = \mathbb{E}[\mathcal{L}(x \mid \mathcal{D}, \Pi(t))] = \int \mathbb{P}(\mathcal{D} \mid \Pi(t))\mathbb{P}(\Pi(t) \mid x)\ d\Pi(t)$$

$\{\Pi(t)\}_{t\geq 0}$ peut être modélisée par un processus de Wright-Fisher et la loi de $\Pi(t)$ dépend donc de t, des fréquences d'haplotypes initiales et des paramètres de recombinaison, mutation, sélection, ..., du modèle de Wright-Fisher. L'instant $t = 0$ correspond généralement à l'apparition d'un nouvel allèle au QTL. Une description plus approfondie des méthodes basées sur cette approche est donnée au chapitre 8.

4.3.2 Méthodes basées sur l'identité par descendance

On dit que deux haplotypes sont identiques par descendance (IBD) s'ils dérivent du même haplotype ancestral sans qu'aucune recombinaison ne soit intervenue. Tout comme les fréquences d'haplotypes dans la population, cette notion peut être utilisée pour modéliser la probabilité d'avoir un type d'allèle au QTL sachant qu'on a l'haplotype h_n aux marqueurs. Supposons par exemple un QTL biallélique où q est l'allèle d'origine et Q un allèle mutant apparu plus récemment. On appelle haplotype ancestral l'haplotype de marqueurs qui était associé à Q au moment où il est apparu. Si la zone délimitée par les marqueurs inclut le QTL et si l'haplotype de marqueurs h_n est IBD avec l'haplotype ancestral, alors il est clair que $\mathbb{P}(Q \mid h_n) = 1$ et $\mathbb{P}(q \mid h_n) = 0$.

Pour localiser des gènes de maladies, McPeak et Strahs (1999) et Morris et al. (2000) ont proposé une méthode de calcul de vraisemblance basée sur cette notion. Ils introduisent un vecteur caché ω de taille L égale au nombre de marqueurs considérés. Chaque coordonnée ω_l peut prendre deux valeurs : A si le marqueur en question est IBD avec l'haplotype ancestral, et N sinon. Ce vecteur est modélisé par deux chaînes de Markov partant de la position du gène de maladie et allant en direction des deux extrémités de la carte. En effet, le statut IBD d'un marqueur ne dépend que du statut IBD du premier marqueur rencontré en direction du gène de maladie, et des éventuelles recombinaisons qui ont pu avoir lieu entre les deux marqueurs.

Soit $\mathcal{I}$ un vecteur contenant l'information IBD de ω ainsi que d'autres informations cachées comme les fréquences alléliques, le temps t depuis l'apparition de l'allèle mutant, ...Considérons, à chaque génération, un modèle de recombinaisons simple tel que celui décrit en 1.2.2. Définissons également un modèle décrivant la relation entre le génotype du géne recherché et le phénotype étudié (Morris et al. (2000) utilisent un modèle

spécifique aux gènes de maladie, mais pour un QTL on pourrait prendre l'un des modèles présentés en 1.4). Il est alors possible d'obtenir une expression explicite de $\mathbb{P}(\mathcal{D} \mid x, \mathcal{I})$ et donc de calculer la vraisemblance par :

$$\mathcal{L}(x \mid \mathcal{D}) = \mathbb{E}[\mathcal{L}(x \mid \mathcal{D}, \mathcal{I})] = \int \mathbb{P}(\mathcal{D} \mid x, \mathcal{I}) \mathbb{P}(\mathcal{I} \mid x) \, d\mathcal{I}$$

En réalité, l'énumération de toutes les valeurs de $\mathcal{I}$ est impossible, et la simulation de cette intégrale par une méthode de Monte Carlo directe ne serait pas efficace. Pour estimer x, McPeak et Strahs (1999) utilisent en fait l'algorithme EM, et Morris et al. (2000) ont développé un algorithme MCMC combiné à un cadre Bayesien. Dans cet algorithme, la distribution simulée par la chaîne de Markov stationnaire est $\mathbb{P}(x, \mathcal{I} \mid \mathcal{D})$, dont on déduit

$$\mathbb{P}(x \mid \mathcal{D}) = \int \mathbb{P}(x, \mathcal{I} \mid \mathcal{D}) \, d\mathcal{I}$$

en mesurant simplement le temps passé par la chaîne dans l'état $\{x\}$. On simule la chaîne de Markov par l'algorithme de Metropolis-Hastings. Quand la chaîne est dans l'état $(x, \mathcal{I})$, la probabilité d'accepter un nouvel état $(x', \mathcal{I}')$ dépend du ratio :

$$\frac{\mathbb{P}(x', \mathcal{I}' \mid \mathcal{D})}{\mathbb{P}(x, \mathcal{I} \mid \mathcal{D})} = \frac{\mathbb{P}(\mathcal{D} \mid x', \mathcal{I}') \pi(x', \mathcal{I}')}{\mathbb{P}(\mathcal{D} \mid x, \mathcal{I}) \pi(x, \mathcal{I})}$$

où $\pi(.)$ est une distribution préalablement choisie. Or on a déjà dit qu'une expression analytique pouvait être obtenue pour $\mathbb{P}(\mathcal{D} \mid x, \mathcal{I})$. Cette méthode a également été adaptée par Pérez-Enciso (2003) pour la localisation de QTL.

Meuwissen et Goddard (2000) ont également développé une méthode basée sur la notion d'IBD. Contrairement aux méthodes évoquées précédemment, elle utilise pour les phénotypes un modèle linéaire mixte où les effets des haplotypes sont aléatoires. Le modèle simplifié s'écrit

$$z = Xh + \epsilon$$

où $z = (z_1, \ldots, z_{N_s})$ est le vecteur des phénotypes, h est un vecteur de dimension J contenant les effets des haplotypes, et X est une matrice d'incidence. ϵ est un vecteur de résidus de matrice de covariance $\sigma_e^2 Id_{N_s}$. La matrice de covariance des effets des haplotypes est $\sigma_h^2 H_x$, où H_x est la matrice IBD : pour $1 \leq j_1, j_2 \leq J$, $(H_x)_{j_1,j_2}$ représente la probabilité que les haplotypes de marqueurs j_1 et j_2 soient IBD au QTL. Cette formulation n'impose aucune contrainte sur le nombre d'allèles au QTL. La matrice IBD dépend de la position x du QTL. Pour chaque position, H_x est estimée par des simulations du processus de Wright-Fisher pour un temps t et une taille de population N choisis arbitrairement. Meuwissen et

Goddard (2001) ont également proposé une expression approchée de H_x calculée à partir d'un modèle de coalescent. Une fois H_x calculée, les autres paramètres et la vraisemblance en x peuvent être estimés par les méthodes standard du modèle linéaire mixte.

4.3.3 Méthodes basées sur l'arbre de coalescence de l'échantillon

La généalogie de l'échantillon d'haplotypes observés est également une information qui, si elle était connue, pourrait permettre de calculer la vraisemblance des observations. Notons $\mathcal{A}$ cette information, dont le contenu exact varie d'une méthode à l'autre mais qui contient toujours au moins :

- L'arbre de coalescence de l'échantillon observé **pour le QTL**.
- Le ou les haplotypes ancestraux associés à l'allèle le plus récent du QTL.
- Certaines informations supplémentaires comme les mutations intervenues le long des branches de l'arbre de coalescence, les haplotypes au niveau des noeuds de l'arbre, ...

Ces informations sont en tout cas assez précises pour que le calcul de $\mathbb{P}(\mathcal{D} \mid x, \mathcal{A})$ puisse être réalisé. Comme précédemment, l'espace dans lequel vit $\mathcal{A}$ est trop grand pour que l'on puisse énumérer toutes les combinaisons. D'autre part, on a vu qu'il était très facile de simuler des arbres de coalescence, mais les méthodes de simulation directe par Monte Carlo seraient peu efficaces ici car pour la plupart des réplicats $\mathcal{A}_i$ la probabilité $\mathbb{P}(\mathcal{D} \mid x, \mathcal{A}_i)$ est très faible. Il faut donc simuler $\mathcal{A}$ conditionnellement à $\mathcal{D}$.

Plusieurs algorithmes MCMC, fonctionnant sur le même principe que celui décrit dans la section précédente, ont donc été développés. La plupart visent à localiser des gènes de maladie dans le cadre d'études cas-contrôle (Graham et Thompson, 1998; Rannala et Reeve, 2001; Morris et al., 2002). L'arbre de coalescence n'est alors reconstruit que pour les individus malades. Plus récemment, Zöllner et Pritchard (2005) ont utilisé ce type d'approches pour la cartographie de QTL, en reconstruisant le coalescent pour tous les individus observés indépendamment de leur phénotype.

Une autre option consiste à inclure dans $\mathcal{A}$ la généalogie des haplotypes entiers au lieu de celle du seul QTL. On peut alors s'appuyer sur les propriétés de l'*ancestral recombination graph* présenté en 2.2.2. Là encore, la taille de l'espace à explorer est immense et il faut trouver des méthodes de simulation efficaces. Pour la cartographie de gènes de maladie ou l'estimation du taux de recombinaison entre deux marqueurs, des méthodes MCMC (Kuhner et al., 2000) ou d'*importance sampling* (Stephens et Donnelly, 2000; Larribe et al., 2002) ont été développées pour calculer la vraisemblance. L'*importance*

sampling consiste à calculer

$$\mathcal{L}(x \mid \mathcal{D}) = \int \mathbb{P}(\mathcal{D} \mid x, \mathcal{A}) \frac{\mathbb{P}(\mathcal{A} \mid x)}{Q(\mathcal{A})} Q(\mathcal{A}) \, d\mathcal{A} \approx \frac{1}{M} \sum_{i=1}^{M} \mathbb{P}(\mathcal{D} \mid x, \mathcal{A}_i) \frac{\mathbb{P}(\mathcal{A}_i \mid x)}{Q(\mathcal{A}_i)}$$

en simulant $\mathcal{A}_1, \ldots, \mathcal{A}_M$ d'après une distribution $Q(.)$ bien choisie. En particulier, $Q(.)$ doit attribuer plus de poids aux arbres $\mathcal{A}$ pour lesquels la probabilité $\mathbb{P}(\mathcal{D} \mid x, \mathcal{A})$ est élevée, afin que les simulations soient efficaces.

4.4 Conclusions

Comme nous venons de le voir au cours de ce chapitre, il existe de nombreux moyens d'aborder le problème de la cartographie fine de gènes par déséquilibre de liaison. La question reste pourtant très ouverte, dans la mesure où beaucoup de méthodes ont été conçues spécifiquement pour les études cas-contrôle, et doivent être repensées dans le cas des QTL. D'autre part aucun consensus ne semble encore se dégager autour d'une méthode particulière, et toutes sont encore perfectibles. Le Tableau 4.1 fait un bilan des principales méthodes déjà existantes pour la cartographie fine de QTL par déséquilibre de liaison, et résume certaines de leurs propriétés.

Dans le développement de nouvelles méthodes, on cherche naturellement à améliorer la qualité des estimations de x, que ce soit en terme d'estimation ponctuelle ou d'intervalles de confiance. Mais d'autres critères nous importent, notamment :

- La capacité d'utiliser conjointement l'information de plusieurs marqueurs. En effet les génotypes des différents marqueurs d'une carte sont généralement très corrélés, et on perd de l'information en les étudiant indépendamment les uns des autres.
- La prise en compte par le modèle de la complexité des phénomènes génétiques. Par exemple, dans les populations animales ou végétales utilisées en agriculture, les individus sont fortement sélectionnés en fonction de leurs phénotypes. Ne pas en tenir compte dans les modèles peut conduire à des erreurs d'estimation.
- La complexité algorithmique. Le volume de données disponibles pour la cartographie est de plus en plus conséquent, tant au niveau du nombre d'individus étudiés que du nombre de marqueurs analysés. Le temps d'exécution des méthodes de cartographie doit donc rester raisonnable dans ce cas.

Comme le montre le Tableau 4.1, beaucoup de méthodes permettent déjà, en théorie, d'utiliser conjointement l'information de tous les marqueurs de la carte. En pratique, la taille de l'espace à explorer augmente exponentiellement avec le nombre de marqueurs,

Propriétés statistiques

Méthode	Modèle de population	Approche	Méthode d'estimation*
Fan et Xiong (2002)	non	-	MV
Jung et al. (2005)	non	-	MV
Tzeng et al. (2006)	non	-	MV
Abdallah et al. (2004)	oui	fréquentiste	MV
Pérez-Enciso (2003)	oui	IBD	B
Meuwissen et Goddard (2000, 2001)	oui	IBD	MV
Zöllner et Pritchard (2005)	oui	coalescence	B

* : Maximum de vraisemblance (MV) ou Bayesien (B)

Propriétés du modèle

Méthode	Nombre de marqueurs	Dominance	Mutations
Fan et Xiong (2002)	2	oui	-
Jung et al. (2005)	illimité	oui	-
Tzeng et al. (2006)	illimité	non	-
Abdallah et al. (2004)	1	oui	non
Pérez-Enciso (2003)	illimité	oui	oui
Meuwissen et Goddard (2000, 2001)	illimité	non	non
Zöllner et Pritchard (2005)	illimité	oui	oui

TAB. 4.1 – Récapitulatif des principales méthodes de cartographie de QTL par déséquilibre de liaison et de leurs caractéristiques

et les méthodes essayant de parcourir cet espace par des algorithmes MCMC sont difficilement applicables. En outre, les méthodes basées sur la reconstruction de la généalogie de l'échantillon sont peu adaptées à des populations sous sélection. En effet bien que le coalescent puisse être étendu pour tenir compte de la sélection (Neuhauser et Krone, 1997; Barton et al., 2004), on perd dans ce cas la propriété essentielle d'indépendance entre la structure des généalogies et la composition en haplotypes de la population. Ceci rend le modèle moins maniable.

Les méthodes que j'ai étudiées durant ma thèse s'inscrivent dans l'approche fréquentiste qui a été introduite en 4.3.1. Comme plusieurs autres, cette approche a l'avantage de s'appuyer sur les modèles classiques de génétique de population pour exploiter au mieux l'information fournie par les haplotypes. Toutefois le modèle utilisé -celui de Wright-Fisher- est assez simple, de sorte qu'on peut espérer régler certaines questions de manière analytique, et éviter ainsi le recours intensif aux simulations. Ce modèle permet en outre d'incorporer assez naturellement des phénomènes de mutation ou de sélection. Nous ver-

rons que les méthodes existantes basées sur cette approche peuvent être améliorées de plusieurs manières : la vraisemblance $\mathcal{L}(x \mid \mathcal{D})$ peut être calculée de manière plus précise grâce à une meilleure prise en compte de la loi des fréquences d'haplotypes (cf partie II) et le nombre de marqueurs utilisés simultanément peut être augmenté (cf partie III).

Références

Abdallah, J., Mangin, B., Goffinet, B., Cierco-Ayrolles, C., et Pérez-Enciso, M. (2004). A comparison between methods for linkage disequilibrium fine mapping of quantitative trait loci. *Genet Res*, *83*, 41-47.

Barton, N., Etheridge, A., et Sturm, A. (2004). Coalescence in a random background. *Amnals of Applied Probabilitiy*, *14*, 754-785.

Boerwinkle, E., Chakraborty, R., et Sing, C. (1986). The use of measured genotype information in the analysis of quantitative phenotypes in man. *Am J Hum Genet*, *50*, 181-194.

Boerwinkle, E., Viscikis, S., Welsh, D., Steinmetz, S., J. Hamash, et Sing, C. (1987). The use of measured genotype information in the analysis of quantitative phenotypes in man. ii. the role of the apolipoprotein e polymorphisms in determining levels, variablility, and covariability of cholesterol, betalipoprotein, and triglycerides in a sample of unrelated individuals. *Am J Med Genet*, *27*, 567-582.

Devlin, B., et Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, *29*, 311-322.

Fan, R., et Xiong, M. (2002). High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. *Eur. J. Hum. Genet.*, *10*, 607-615.

Graham, J., et Thompson, E. (1998). Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am J Hum Genet*, *63*, 1517-1530.

Hästbacka, J., Chapelle, A. de la, Kaitila, I., Sistonen, P., Weaver, A., et Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations : diastrophic dysphasia in finland. *Nat Genet*, *2*, 204-211.

Jorde, L. (1995). Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet*, *52*, 11-14.

Jung, J., Fan, R., et Jin, L. (2005). Combined linkage and association mapping of quantitative trait loci by multiple markers. *Genetics*, *170*, 881-898.

Kuhner, M., Yamato, J., et Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics*, *156*, 1393-1401.

Larribe, F., Lessard, S., et Schork, N. (2002). Gene mapping via the ancestral recombination graph. *Theor. Pop. Biol.*, *62*, 215-229.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., et al. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.*, *78*, 437-450.

McPeak, M., et Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine scale genetic mapping. *Am J Hum Genet*, *65*, 858-875.

Meuwissen, T., et Goddard, M. (2000). Fine mapping of quantitative trait loci using linkage disequilibrium with closely linked marker loci. *Genetics*, *155*, 421-430.

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

Molitor, J., Marjoram, P., et Thomas, D. (2003). Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am. J. Hum. Genet.*, *73*, 1368-1384.

Morris, A., Whittaker, J., et Balding, D. (2000). Bayesian fine-scale mapping of disease loci by hidden markov models. *Am J Hum Genet*, *67*, 155-169.

Morris, A., Whittaker, J., et Balding, D. (2002). Fine-scale mapping of disaese loci via shattered coalescent modelling of genealogies. *Am J Hum Genet*, *76*, 686-707.

Neuhauser, C., et Krone, S. (1997). The genealogy of samples in models with selection. *Genetics*, *145*, 519-534.

Nielsen, D., et Weir, B. (1999). A classical setting for associations between markers and loci affecting quantitative traits. *Genet. Res*, *74*, 271-277.

Niu, T. (2004). Algorithms for inferring haplotypes. *Genetic Epidemiology*, *27*, 334-347.

Pérez-Enciso, M. (2003). Fine mapping of complex trait genes combinig pedigree and linkage disequilibrium information : a bayesian unified framework. *Genetics*, *163*, 1497-1510.

Rannala, B., et Reeve, J. (2001). High resolution multipoint linkage disequilibrium mapping in the context of a human genome sequence. *Am J Hum Genet*, *69*, 159-178.

Stephens, M., et Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Statist. Soc.*, *62*, 605-655.

Tzeng, J., Wang, C., Kao, J., et Hsiao, C. (2006). Regression-based association analysis with clustered haplotypes through use of genotypes. *Am. J. Human. Genet.*, *78*, 231-242.

Waldron, E., Whittaker, J., et Balding, D. (2006). Fine mapping of disease genes via haplotype clustering. *Genetic Epidemiology*, *30*, 70-179.

Zöllner, S., et Pritchard, J. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, *169*, 1071-1092.

Deuxième partie

Distribution des fréquences d'haplotypes sous un modèle de Wright-Fisher à deux loci

Chapter 5

Article : Probabilty distribution of haplotype frequencies under the two-locus Wright-Fisher model by diffusion approximation

Auteurs

Simon Boitard et Patrice Loisel (INRA Montpellier, Laboratoire d'Analyse des Systèmes et Biométrie)

Statut

En relecture à *Theoretical Population Biology*

Résumé en français

Caractériser la loi des fréquences d'haplotypes dans une population pour des systèmes de plusieurs loci liés est essentiel pour la cartographie de gènes par déséquilibre de liaison. Cela peut notamment permettre de mieux comprendre l'influence de paramètres tels que la sélection, la mutation, le temps d'évolution, ..., sur les niveaux de déséquilibre de liaison observés à travers les différentes mesures. Nous avons également vu au chapitre 4 que pour certaines méthodes de cartographie la vraisemblance dépendait de la loi des fréquences d'haplotypes.

Il est généralement plus facile d'étudier le modèle de diffusion que le modèle de Wright-Fisher lui-même. C'est ce que nous faisons dans cet article pour le modèle à deux loci présenté en 5.2.1. Le générateur de la diffusion obtenue comme limite de ce modèle a été donnée par Ethier et Nagylaki (1989). Nous considérons deux loci biallélliques, et les fréquences d'haplotypes à l'instant τ sont donc représentées par le vecteur $Y(\tau) = (Y_{1,1}(\tau), Y_{1,2}(\tau), Y_{2,1}(\tau))$ qui est à valeurs dans l'espace

$$\mathbb{F}_3 = \{y \in [0,1]^3, y_1 + y_2 + y_3 \leq 1\}$$

Notre objectif est de déterminer, pour des fréquences initiales $y^0 \in \mathbb{F}_3$ et un temps $\tau \in \mathbb{R}$ fixés, la loi de $Y(\tau)$. Ceci revient à calculer, pour tout $y \in \mathbb{F}_3$, la densité de transition $f(y^0, y, \tau)$ du processus de diffusion. On sait que pour y et y^0 à l'intérieur du domaine $\mathbb{F}_3$, cette quantité vérifie l'équation projective de Kolmogorov

$$\frac{\partial f}{\partial \tau} = Lf$$

où L est le générateur du processus de diffusion.

Nous décrivons dans l'article la méthode numérique que nous avons développée pour calculer la densité de transition à partir de cette équation et de la condition initiale $Y(0) = y^0$. Elle est basée sur un shéma de différences finies et des conditions aux bords approchées. Nous utilisons ensuite certaines propriétés connues de la loi (espérance, loi des fréquences alléliques marginales ...) pour tester la précision de la solution obtenue par notre méthode en fonction des valeurs de τ, y^0 et du taux de recombinaison. Nous comparons également la complexité algorithmique de notre méthode avec celle d'autres méthodes permettant d'aboutir au même résultat, comme les simulations de Monte Carlo par exemple.

Nous en concluons que notre méthode est particulièrement intéressante pour des temps d'évolution courts, des taux de recombinaisons faibles, et des fréquences d'haplotypes

initiales pas trop proches de 0 ou 1. Dans ce cas l'erreur commise est très faible. D'autre part, pour des populations de taille importante, le temps de calcul est largement inférieur à celui des autres méthodes considérées. Notre méthode semble donc un outil intéressant pour l'étude des fréquences d'haplotypes, et notamment du déséquilibre de liaison, en régime transitoire.

Abstract

The probability distribution of haplotype frequencies in a population, and the way it is influenced by genetical forces such as recombination, selection, random drift ... is a question of fundamental interest in population genetics. For large populations, the distribution of haplotype frequencies for two linked loci under the classical Wright-Fisher model is almost impossible to compute because of numerical reasons. However the Wright-Fisher process can in such cases be approximated by a diffusion process and the transition density can then be deduced from the Kolmogorov equations. As no exact solution has been found for these equations, we developed a numerical method based on finite differences to solve them. It applies to transient states and models including selection or mutations. We show by several tests that this method is accurate for computing the conditional joint density of haplotype frequencies given that no haplotype has been lost. We also prove that it is far less time consuming than other methods such as Monte Carlo simulations.

Keywords

Wright-Fisher model, Diffusion processes, Kolmogorov forward equation, Finite difference scheme

5.1 Introduction

The mathematical study of the dynamics of allele frequencies within species is a central question in population genetics. In this context, the model of random genetic drift introduced by Wright and Fisher in the early 1920's has played a major role, and many well accepted results in genetics are based on it. We consider here a version of this model for two linked loci, as introduced by Ethier and Nagylaki (1980). Let $\Pi(t)$ be the vector of the haplotype frequencies in the population at time t. We are interested in solving two different but related problems, respectively denoted forward and backward.

The forward problem is rather theoretical: given the composition $\Pi(0)$ of the population at time 0, we would like to know the probability distribution of the composition, $\Pi(t)$, t generations later. We find typical examples of this problem in the classical studies about the fixation rate of an allele (Karlin and McGregor, 1968) or the properties of linkage disequilibrium measures such as

$$D_{i_1,i_2}(t) = \Pi_{i_1,i_2}(t) - \Pi_{i_1,.}(t)\Pi_{.,i_2}(t)$$

or

$$r^2_{i_1,i_2}(t) = \frac{D^2_{i_1,i_2}(t)}{\Pi_{i_1,.}(t)(1-\Pi_{i_1,.}(t))\Pi_{.,i_2}(t)(1-\Pi_{.,i_2}(t))} \tag{5.1}$$

where $\Pi_{i_1,i_2}(t)$, $\Pi_{i_1,.}(t)$ and $\Pi_{.,i_2}(t)$ respectively denote the frequency of haplotype (i_1, i_2), of allele i_1 at first locus and of allele i_2 at second locus in the population. As illustrated in several reviews about these measures (Devlin and Risch, 1995; Cierco-Ayrolles et al., 2004), most studies focus on the expected value and variance instead of on the whole distribution. They also require quite strong hypotheses, essentially neutral models at stationarity (Ohta and Kimura, 1969; Sved, 1971; Griffiths, 1981; Hudson, 1985). Some transient properties of the two-locus model were reported by Litler (1973), but his results were biased by all the trajectories leading to fixation at one of the two loci. It seems more relevant to study the distribution of haplotype frequencies given that no locus fixed. This is the approach recently used by Mano (2005), who derived a formula for the conditional expected value of the frequency of one haplotype under a two-locus neutral model. Simplifications of the general problem are necessary because of the difficulty to obtain exact results about the transition density of the two-locus model. An alternative solution, which can apply to very general models, is to simulate forward the Wright-Fisher process in order to study its properties. This approach is widely used in genetics. Unfortunately it is quite inefficient for determining the whole transition density, due to the size of the state space.

The backward problem consists reversely in determining the probability distribution of $\Pi(0)$ given $\Pi(t)$. It is more related to real applications because we generally observe populations at present and try to infer their history. If we know for example that an important demographic event (migration, bottleneck ...) occurred approximately t generations ago, it can be useful to have an idea of the most probable haplotype frequencies at that time. If we then select the density value of the most probable initial haplotype composition, we get a likelihood value for the genetic parameters (recombination rate, selection rate, mutation rate ...) used in the model. This objective of parameter inference in a multi-locus context is central in population genetics. During the last decade, most methods have been based on the ancestral recombination graph or approximations of it (Nordborg, 2001).

Under the Wright-Fisher model, the haplotype composition $\Pi(t)$ is a discrete Markov chain evolving in a discrete state space of size $\frac{(2N)^{C-1}}{(C-1)!}$, where C is the number of possible haplotypes (2^L for L bi-allelic loci for instance) and N is the population size. The computation of the probability distribution of $\Pi(t)$ at any time t, whereas straightforward in theory, becomes quickly intractable for large population sizes. The first step is thus to approximate $\Pi(t)$ by a diffusion process with continuous time and continuous state space. The idea was introduced by Fisher (1922) and Wright (1931), and was later greatly extended by Kimura (1964). It provides partial differential equations for most quantities related with $\Pi(t)$, thanks to Kolmogorov theorems. In the one-locus model, some of these equations were solved and nice results were thus obtained, as illustrated by Karlin and Taylor (1981) or Crow and Kimura (1970). In the two-locus model, the diffusion is multidimensional and Kolmogorov theorems lead to partial differential equations that generally cannot be solved in analytic form, especially during the transient period of the process.

To overcome this problem, we developed a numerical method based on finite differences to solve this kind of equations. The Kolmogorov forward equation and the Kolmogorov backward equation can then be used to solve the forward problem and the backward problem respectively. In this paper we focus on the forward problem, and study an approximation of the transition density of the haplotype frequency vector $\Pi(t)$ with two bi-allelic loci. We condition the density on the fact that all haplotypes are still present at time t. Our method applies to any transient state, and to models including mutations between alleles or weak directional selection.

We begin with a short definition of the Wright-Fisher models with one and two loci and their diffusion limits in Section 5.2. These notions, even the ones about the one-locus model, will be useful in what follows. We then present the numerical scheme in Section 5.3. Section 5.4 is devoted to the accuracy of this scheme for computing the transition

density of haplotype frequencies under the diffusion model. We then illustrate the interest of our method in Section 5.5 by an example about the linkage disequilibrium measure $r^2_{i,j}$. We also prove in Section 5.6 that the results obtained by this method could hardly be obtained by direct methods based on the Wright-Fisher model. Finally in Section 5.7 we discuss the advantages and drawbacks of our approach in terms of accuracy and domain of validity.

5.2 Models

5.2.1 Wright-Fisher models

The Wright-Fisher model is the basic model for haplotype frequency evolution in isolated randomly mating diploid populations of finite size. Generations are non-overlapping and the population size N is held constant. Each generation is determined through a multinomial sampling of the previous one, where sampling probabilities depend on diverse forces such as selection or mutation. Here we present the case with no selection or mutation, both for one and two loci. One-locus models with selection and mutation can be found in chapter 10 of (Ethier and Kurtz, 1986), and two-locus models are presented in (Ethier and Nagylaki, 1980).

The one-locus model

Considering the evolution of one locus with I alleles, let $X(t)$ be the number of copies of allele 1 in the population at time $t \in \mathbb{N}$. The Wright-Fisher model postulates that the process $\{X(t)\}_{t\in\mathbb{N}}$ is a Markov chain with state space $\{0,\dots,2N\}$ and binomial transition probability

$$\mathbb{P}(X(t+1)=z \mid X(t)=x) = \binom{2N}{z}(\frac{x}{2N})^z(1-\frac{x}{2N})^{2N-z}$$

The two-locus model

The two-locus model is based on the notion of haplotypes, that are the pairs of alleles (i_1,i_2) where i_1 refers to the first locus and i_2 to the second one. We denote $X_{i_1,i_2}(t)$ the number of haplotypes (i_1,i_2) in the population at generation t ($i_1 = 1,\dots,I_1$ and $i_2 = 1,\dots,I_2$) and define $X(t) = (X_{1,1}(t),\dots,X_{I_1,I_2}(t))$ the vector of haplotype counts. The distribution of $X(t+1)$ given $X(t) = x = (x_{i_1,i_2})_{1\le i_1\le I_1, 1\le i_2\le I_2}$ is the multinomial

distribution $\mathcal{M}(2N, p_{1,1} \dots p_{I_1,I_2})$ with parameters

$$p_{i_1,i_2} = (1-c)\frac{x_{i_1,i_2}}{2N} + c\frac{x_{i_1,.}}{2N}\frac{x_{.,i_2}}{2N} \qquad i_1 = 1, \dots, I_1,\ i_2 = 1, \dots, I_2 \tag{5.2}$$

where $x_{i_1,.} = \sum_{i_2=1}^{I_2} x_{i_1,i_2}$ and $x_{.,i_2} = \sum_{i_1=1}^{I_1} x_{i_1,i_2}$. $c \in [0, 1/2]$ is the recombination rate between the two loci. If $c = 0$, the two loci are totally linked and haplotype counts can be studied exactly as allele counts in the one-locus model. Otherwise the distribution of $X_{i_1,i_2}(t+1)$ depends not only on $X_{i_1,i_2}(t)$ but on all other haplotype counts at time t as well. Due to the relation $\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} X_{i_1,i_2}(t) = 2N$, we can skip the last component and study the vector

$$(X_{1,1}(t), \cdots, X_{I_1,I_2-1}(t))$$

It is a Markov chain evolving in the state space

$$\mathbb{F}_{I_1 I_2 - 1} = \{x \in \{0, \cdots, 2N\}^{I_1 I_2 - 1}, x_1 + \cdots + x_{I_1 I_2 - 1} \leq 2N\}$$

From this point we use the notation $X(t)$ for this vector and no longer for the one with all components.

5.2.2 Diffusion approximations and Kolmogorov equations

For large populations, the use of the transition matrix of the one-locus or two-locus Markov chain to compute the distribution is not efficient because of the state space size. However, we give in this section two theorems proving that as $N \to +\infty$, Wright-Fisher processes can be approximated by diffusion processes which are easier to handle with.

One-locus model

Let $\{X^N(t)\}_{t\in\mathbb{N}}$ follow the one-locus Wright-Fisher model described in Section 5.2.1. The superscript N is added here for the sake of clarity. Let us define, for all $\tau \geq 0$, $Y^N(\tau) = \Pi^N([2N\tau]) = X^N([2N\tau])/(2N)$, where [ ] denotes the integer part. For any compact set E, we also introduce the space $D_E[0,\infty)$ of all càd-làg functions $f : [0,\infty) \to E$ endowed with the Skorohod metric (Billingsley, 1968).

Theorem 1 *As $N \to \infty$, the scaled process $\{Y^N(\tau)\}_{\tau\geq 0}$ converges weakly in $D_{[0,1]}[0,\infty)$ to a Markovian diffusion process $\{Y(\tau)\}_{\tau\geq 0}$ with state space $[0,1]$ and generator $Lf(y) = \frac{1}{2}y(1-y)\frac{\partial^2 f}{\partial y^2}(y)$, $f \in C^2([0,1])$.*

The proof of this theorem can be found in (Ethier and Kurtz, 1986).

For $0 < y < 1$, the infinitesimal variance $a^2(y) = y(1-y)$ is strictly non negative, independent of time and twice continuous with bounded derivatives. Thus for $0 < y^0, y < 1$, the transition function $P(y^0, y, \tau)$ of $\{Y(\tau)\}_{\tau \geq 0}$ is strongly continuous with respect to Lebesgue's measure (Stroock and Varadhan, 1997) and its density $f(y^0, y, \tau)$ satisfies the Kolmogorov forward equation

$$\begin{aligned} \frac{\partial f(y^0, y, \tau)}{\partial \tau} &= L^* f(y^0, y, \tau) \\ &= \frac{1}{2} \frac{\partial^2 [y(1-y) f(y^0, y, \tau)]}{\partial y^2} \qquad \tau > 0,\ 0 < y^0, y < 1 \end{aligned} \tag{5.3}$$

L^* denotes the adjoint infinitesimal generator of $\{Y(\tau)\}$. The solution of this equation is (Kimura, 1955b)

$$f(y^0, y, \tau) = y^0(1-y^0) \sum_{n=0}^{+\infty} C_n e^{-\frac{1}{2}(n+1)(n+2)\tau} P_n^{1,1}(1-2y^0) P_n^{1,1}(1-2y) \tag{5.4}$$

where $C_n = (2n+3)(n+2)/(n+1)$ and $\{P_n^{1,1}\}_{n \in \mathbb{N}}$ are the Jacobi polynomials (Erdélyi, 1953). $f(y^0, ., \tau)$ is not a probability density because $\int_0^1 f(y^0, y, \tau)\ dy < 1$ (this comes from the non zero probability of allele loss $Y(\tau) = 0$ or 1). In this paper we consider the conditional density given $0 < Y(\tau) < 1$, dividing $f(y^0, y, \tau)$ by $\int_0^1 f(y^0, y, \tau)\ dy$. It is a probability measure.

Kimura (1955a) also found the joint transition density of allele frequencies for a model with three alleles, and Littler and Fackerell (1975) extended it to an arbitrary number of alleles. Griffiths (1979) provided an expansion in orthogonal polynomials for the transition density in a multi-allele diffusion model with a particular model of mutations between alleles. For models including selection, several approximations were proposed. Kimura (1955c) solved the Kolmogorov forward equation for a model with varying selection using an approximation of the infinitesimal drift. Tier and Keller (1978) presented a general ray method for obtaining short time asymptotic solutions of diffusion equations and applied it to several genetical models with one locus or two independent loci. More recently, Barbour et al. (2000) provided an exact but non explicit expansion for the transition function of a model including mutation and selection. This expansion depends on the distribution of a certain birth-and-death process starting at "infinity".

Two-locus model

Let $\{X^N(t)\}_{t \in \mathbb{N}}$ follow the two-locus Wright-Fisher model described in Section 5.2.1 and let us define, for all $\tau \geq 0$, $Y^N(\tau) = X^N([2N\tau])/(2N)$. For the sake of clarity, we now

index haplotypes by $j = 1 \dots J$, with $J = I_1 I_2 - 1$, and we introduce the set $\mathbb{F}_J = \{y \in [0,1]^J, y_1 + \dots + y_J \leq 1\}$.

Theorem 2 *Assume that $\rho = \lim_{N \to +\infty} 2Nc$ exists. Then, as $N \to \infty$, the scaled process $\{Y^N(\tau)\}_{\tau \geq 0}$ converges weakly in $D_{\mathbb{F}_J}[0,\infty)$ to a Markovian diffusion process $\{Y(\tau)\}_{\tau \geq 0}$ with state space $\mathbb{F}_J$ and generator*

$$Lf(y) = \sum_{j=1}^{J} b_j(y) \frac{\partial f}{\partial y_j}(y) + \frac{1}{2} \sum_{j_1, j_2 = 1}^{J} a^2_{j_1, j_2}(y) \frac{\partial^2 f}{\partial y_{j_1} \partial y_{j_2}}(y)$$

where $b_j(y) = \rho(-y_j + y_{h_j(1),.}\ y_{.,h_j(2)})$ and $a^2_{j_1,j_2}(y) = y_{j_1}(\delta_{j_1;j_2} - y_{j_2})$.

$h_j(1)$ and $h_j(2)$ respectively denote the alleles at first and second locus of haplotype j, and $\delta_{j_1;j_2}$ is the Kronecker symbol equal to 1 if $j_1 = j_2$ and 0 otherwise. The proof can be found in (Ethier and Nagylaki, 1989) and holds true for more complex models as well.

This diffusion process has the same good properties as the one obtained with one locus. Thus for y^0 and y in the interior of $\mathbb{F}_J$, denoted $\mathring{\mathbb{F}}_J$, and $\tau > 0$ the transition density $f(y^0, y, \tau)$ exists and satisfies the Kolmogorov Forward Equation

$$\begin{aligned} \frac{\partial f(y^0, y, \tau)}{\partial \tau} &= L^* f(y^0, y, \tau) \\ &= -\sum_{j=1}^{J} \frac{\partial (b_j(y) f(y^0, y, \tau))}{\partial y_j} + \frac{1}{2} \sum_{j_1, j_2 = 1}^{J} \frac{\partial^2 (a^2_{j_1, j_2}(y) f(y^0, y, \tau))}{\partial y_{j_1} \partial y_{j_2}} \end{aligned}$$

No analytic exact solution for this equation has ever been found. In the next section we present a numerical framework that we built up for computing the conditional density

$$g(y^0, y, \tau) = \frac{f(y^0, y, \tau)}{\mathbb{P}(Y(\tau) \in \mathring{\mathbb{F}}_J \mid Y(0) = y^0)} \quad \tau > 0,\ 0 < y^0, y < 1$$

In what follows, we generally omit the reference to y^0 and denote $f(y, \tau)$ the transition density and $g(y, \tau)$ the conditional transition density.

5.3 Numerical solution for the two-locus model

We developed our method in the case of two bi-allelic loci, which lead to consider $J = 3$ different haplotypes. The vector of interest, $Y(\tau)$, is of dimension three and the forward

equation to be solved, once developed, is

$$\frac{\partial f(y,\tau)}{\partial \tau} = (\rho - 6) f(y,\tau) + \sum_{j=1}^{3} \breve{b}_j(y) \frac{\partial f(y,\tau)}{\partial y_j} + \frac{1}{2} \sum_{j_1,j_2=1}^{3} a^2_{j_1,j_2}(y) \frac{\partial^2 f(y,\tau)}{\partial y_{j_1} \partial y_{j_2}} \tag{5.5}$$

with

$$a^2(y) = \begin{pmatrix} y_1(1-y_1) & -y_1 y_2 & -y_1 y_3 \\ -y_1 y_2 & y_2(1-y_2) & -y_2 y_3 \\ -y_1 y_3 & -y_2 y_3 & y_3(1-y_3) \end{pmatrix}$$

and

$$\breve{b}(y) = \begin{pmatrix} 1 - 4y_1 + \rho(y_1 - (y_1+y_2)(y_1+y_3)) \\ 1 - 4y_2 + \rho(y_2 - (y_1+y_2)(1-y_1-y_3)) \\ 1 - 4y_3 + \rho(y_3 - (1-y_1-y_2)(y_1+y_3)) \end{pmatrix}$$

The initial condition is $f(y,0) = \delta_{y^0}(y)$, where δ_{y^0} is the Dirac delta function in y^0. So far we don't specify the boundary conditions. We will come back to this problem later.

5.3.1 Operator discretization

Equations like (5.5) where derivatives with respect to time and space are clearly separated are called evolution equations. A natural approach to solve this kind of problems is the finite difference method (Lucquin and Pironneau, 1995). We discretize $[0,\tau]$ with $N_\tau + 1$ equally spaced points and each dimension of $[0,1]^3$ with $N_y + 1$ equally spaced points. For $n = 0, \ldots, N_\tau$, we define the three-dimensional matrix $\tilde{f}^{(n)} \in \mathcal{M}_{\{0,\ldots,N_y\}^3}(\mathbb{R})$ by $\tilde{f}^{(n)}_{k,l,m} = \tilde{f}(k\Delta_y, l\Delta_y, m\Delta_y, n\Delta_\tau)$, where $\delta_y = 1/N_y$ and $\Delta_\tau = \tau/N_\tau$. $\tilde{f}^{(0)}$ approximates $f(.,0)$, i.e. $\tilde{f}^{(0)}(k,l,m)$ is 1 for the index (k,l,m) that corresponds to y^0 and 0 otherwise. $\tilde{f}^{(N_\tau)}$ is the approximate solution of (5.5). It is given by N_τ iterations of the relation

$$\frac{\tilde{f}^{(n+1)} - \tilde{f}^{(n)}}{\delta_\tau} = \tilde{L}^* \tilde{f}^{(n)} \tag{5.6}$$

where

$$\tilde{L}^* = (c-6) Id + \sum_{j=1}^{3} \tilde{b}_j \otimes D^j(\tilde{b}_j) + \sum_{j_1,j_2=1}^{3} \tilde{a}^2_{j_1,j_2} \otimes D^{j_1,j_2}$$

is a discrete version of the operator L^*. In this formula:

- Id is the identity operator on $\mathcal{M}_{\{0,\ldots,N_y\}^3}(\mathbb{R})$.
- $\tilde{b}_j \in \mathcal{M}_{\{0,\ldots,N_y\}^3}(\mathbb{R})$ with $(\tilde{b}_j)_{\ k,l,m} = \breve{b}_j(k\Delta_y, l\Delta_y, m\Delta_y)$.
- $\tilde{a}^2_{j_1,j_2} \in \mathcal{M}_{\{0,\ldots,N_y\}^3}(\mathbb{R})$ with $(\tilde{a}^2_{j_1,j_2})_{\ k,l,m} = a^2_{j_1,j_2}(k\Delta_y, l\Delta_y, m\Delta_y)$.

- $\otimes$ denotes the term by term product for matrix.
- $D^j(\tilde{b}_j)$ and D^{j_1,j_2} are operators on $\mathcal{M}_{\{0,\dots,N_y\}^3}(\mathbb{R})$ such that for any $\lambda \in \mathcal{M}_{\{0,\dots,N_y\}^3}(\mathbb{R})$
  - $(D^1(\tilde{b}_1)\lambda)_{k,l,m} = \begin{cases} \frac{\lambda_{k+1,l,m}-\lambda_{k,l,m}}{\Delta_y} & \text{if} \quad (\tilde{b}_1)_{k,l,m} \geq 0 \\ \frac{\lambda_{k,l,m}-\lambda_{k-1,l,m}}{\Delta_y} & \text{if} \quad (\tilde{b}_1)_{k,l,m} \leq 0 \end{cases}$
  - $(D^{1,1}\lambda)_{k,l,m} = \frac{\lambda_{k+1,l,m}-2\lambda_{k,l,m}+\lambda_{k-1,l,m}}{\Delta_y^2}$
  - $(D^{1,2}\lambda)_{k,l,m} = \frac{\lambda_{k+1,l+1,m}-\lambda_{k+1,l-1,m}-\lambda_{k-1,l+1,m}+\lambda_{k-1,l-1,m}}{4\Delta_y^2}$
  - Definitions of $D^2(b_2)$, $D^3(b_3)$, $D^{2,2}$, $D^{3,3}$ and D^{j_1,j_2} with $j_1 \neq j_2$ are similar.

We thus approximate first order derivatives with an explicit Euler non-centered scheme and second order derivatives with an explicit Euler centered scheme. This discretization method is very classical. It leads to direct implementation, stability is proved for one dimensional schemes with following stability conditions : $\Delta_\tau \leq \Delta_y/v$ for convection equations (with v velocity) and $\Delta_\tau \leq \Delta_y^2/(2\nu)$ for diffusion equations (with ν viscosity) (Lucquin and Pironneau, 1995). We generalize to multi dimensional schemes by choosing $\Delta_\tau = C\min(\Delta_y/||b(.)||, \Delta_y^2/(2||a^2(.)||)$. Since $||a^2(.)|| \leq 1/2$ the choice of $\Delta_\tau = \Delta_y^2$ is suitable. Numerical results confirm this choice. The theoretical error caused by this numerical scheme is of order $\mathcal{O}(\Delta_y + \Delta_\tau) = \mathcal{O}(\Delta_y)$. Other discretizations of L^*, as the characteristics method, were investigated but didn't provide better results in terms of precision (results not shown).

After the N_τ iterations, $\tilde{f}^{(N_\tau)}$ is finally divided by $\int_{\mathbb{F}_J} \tilde{f}^{(N_\tau)}(y)\, dy$ to obtain the approximate conditional density $\tilde{g}$. The computation of the integral is performed by the trapeze method.

5.3.2 Boundary conditions

Kolmogorov equation (5.5) only applies on the interior of $\mathbb{F}_3$ so boundary values of $\tilde{f}$ cannot be computed using (5.6). This is the case for points y such that $y_1 = 0$, $y_2 = 0$, $y_3 = 0$ or $y_1 + y_2 + y_3 = 1$. We are actually not interested in the density values at these points -because we eventually focus on the interior density- but we need them to evaluate the discrete operator L^* at each step of the algorithm. Generally, evolution equations are solved with Dirichlet or Von Neumann boundary conditions, stating respectively that f vanishes or that its first derivative in a given direction vanishes. However, when considering the solution with one locus, for which we know the analytic solution (5.4), it appears that none of these conditions holds.

Determining the exact equations for the boundary points would be extremely difficult. Part of the problem is the diversity of boundary natures (corners, edges and faces) and the fact that $Y(t)$ can have different behaviors depending on the boundary. Indeed, the loss of one allele (as $Y_{1,1} = Y_{1,2} = 0$) is non reversible, which with the terminology of the Feller boundary classification is characteristic of an *exit boundary*. On the contrary, a lost haplotype (as $Y_{1,1} = 0$) can be recovered later by recombination, and this kind of behavior corresponds to a *regular boundary*. More details about boundary classifications can be found in (Karlin and Taylor, 1981).

Consequently, we have chosen a pragmatic and approximate solution consisting in vanishing the second order derivative of $\tilde{f}$ in a direction orthogonal to the boundary. The value at each point of the boundary can thus be deduced from the one of two close points of the interior. This ad hoc method was tested using several accuracy criteria and turned out to be efficient, more than Dirichlet or Von Neuman conditions for example. We present some of these tests in Section 5.4.

5.4 Accuracy of the approximation

In order to compute the probability distribution of haplotype frequencies under the diffusion model, we discretize the Kolmogorov forward equation and use approximate boundary conditions. The amount of error created by this numerical scheme is evaluated in this section for several values of τ, ρ and y^0. As a control, we first consider a one-locus neutral model, for which we are able to compute the exact diffusion density. We expect the results found in this case to hold in the two-locus case as well. Then we try to investigate directly the error made in the two-locus neutral model by studying the particular case of $\rho = 0$ or Monte Carlo simulations.

5.4.1 Numerical solution versus exact solution for one locus

In theory, the error due to the numerical solving of Kolmogorov forward equation (5.3) can be easily controlled through the discretization step Δ_y since the precision of the solution is of order $\mathcal{O}(\Delta_y)$. However the particular boundary conditions we chose could affect this expected performance. Consequently we implemented for (5.3) a numerical scheme similar to the one described in Section 5.3 (see appendix) and compared its result $\tilde{g}$ with the exact solution g. This exact solution can be computed from equation (5.4). Except for very small values of τ which we don't need to consider, the 30 first terms of the series suffice to get a very high accuracy. For the comparison we used the Hellinger's distance

between probability densities

$$d(g,\tilde{g}) = \int_0^1 (\sqrt{g(y)} - \sqrt{\tilde{g}(y)})^2 dy$$

Results are shown in Table 1. We used several values of time τ (from 0.01 to 2) and of initial allele frequency y^0 (from 0.01 to 0.5) leading to different density shapes, and we took a constant discretization step $\Delta_y = 1/100$. The accuracy was always very high. It increased as τ increased, which might come from the fact that the boundary condition $\frac{\partial^2 f}{\partial y^2} = 0$ is satisfied as τ tends to infinity. Unless τ is very small, the coefficients $e^{-\frac{1}{2}(n+1)(n+2)\tau}$ in equation (5.4) actually imply that f is a low-order polynomial in y. Besides, the accuracy of our solution also increased as y^0 increased up to 0.5. This seems natural because the density values near the boundaries, and therefore the consequences of the approximate boundary conditions, are larger for small values of y^0.

5.4.2 Numerical solution versus exact solution for two loci

We now come back to the two-locus model. We remind that it consists in studying the conditional probability density g of the diffusion process

$$Y(\tau) = (Y_1(\tau), Y_2(\tau), Y_3(\tau)) = (Y_{(1,1)}(\tau), Y_{(1,2)}(\tau), Y_{(2,1)}(\tau))$$

given $Y(\tau) \in \mathring{\mathbb{F}}_3$. The numerical scheme described in Section 5.3 provides an approximation $\tilde{g}$ of this density. For $\rho = 0$ (totally linked loci), the exact density g is known and we can directly compare the two functions. In general, the two-locus solution is unknown so we estimate it using Monte Carlo simulations and compare our approximation $\tilde{g}$ with this estimation.

Totally linked loci

With no recombination ($\rho = 0$), the two-locus model with bi-allelic loci can be described as a one-locus model with four alleles. These alleles are the haplotypes of the two-locus model. The joint conditional density g of haplotype frequencies can thus be deduced from the series expansion in (Griffiths, 1979). For very small values of τ such as $\tau = 0.01$, the convergence of this expansion is very slow and the result cannot be used in practice. For $\tau \geq 0.1$, the convergence was quickly achieved and we could compare our approximation $\tilde{g}$ with the series expansion g for several values of time τ and initial haplotype frequencies y^0. The results given in Table 2 show that the accuracy was generally very high, except

for the case where $\tau = 0.1$ and $y^0 = (0.10, 0.01, 0.30)$. As in the one-locus problem of Section 5.4.1, the accuracy was lower for small evolution times and small initial haplotype frequencies. The case of $y^0 = (0.25, 0.25, 0.25)$, not shown in the table, provided even better results.

To illustrate the good accuracy of our approximation, we also computed the marginal density g_2 of the frequency of haplotye $(1, 2)$ by

$$g_2(y_2) = \int_0^{1-y_2} dy_1 \int_0^{1-y_1-y_2} g(y_1, y_2, y_3) dy_3$$

We similarly obtained, from $\tilde{g}$, the approximation $\tilde{g}_2$ of this quantity. We did it for $\tau = 0.1$ and for several values of ρ from 0 to 10. As observed in Figure 5.1, the approximate curve with $\rho = 0$ is very similar to the exact curve. As the recombination rate increases, the approximate density curve is shifted toward large values of y_2. This is natural because the initial frequency of haplotype $(1, 2)$ ($y_2^0 = 0.05$) was lower than the one expected under equilibrium ($(y_1^0 + y_2^0)(1 - y_1^0 - y_3^0) = 0.12$).

Monte Carlo simulations

For $\rho \neq 0$, the exact joint density g of haplotype frequencies is unknown and the quality of our approximation $\tilde{g}$ can thus not be evaluated directly. We can however estimate the true density g using Monte-Carlo simulations of the two-locus diffusion model, or equivalently of the two-locus Wright-Fisher model for a sufficiently large population size N. We chose the latter option. Simulating the evolution of haplotype frequencies under a two-locus Wright-Fisher model is straightforward : it consists in sampling with replacement $2N$ haplotypes at each generation, where the sampling probabilities are given in equation (5.2).

Estimating a density in dimension three with a small discretization step on each direction would actually require a huge number of simulation replicates. To speed up the computations and obtain more reliable estimates, we thus focused on two densities in dimension one. The first one was the marginal density g_P of

$$P(\tau) = Y_{1,1}(\tau) + Y_{1,2}(\tau)$$

which is the frequency of allele 1 at first locus. The second one was the marginal density g_2 of the frequency of haplotype $(1, 2)$. We simulated the independent replicates $\Pi^{(r)}(.)$, $r = 1, \ldots, R$, rejected the ones leading to the loss of at least one haplotype and computed

the Monte Carlo estimates

$$\begin{aligned}\hat{g}_P(k\Delta_y) &= \frac{1}{\Delta_y}\frac{\sum_{r=1}^{R}\mathbb{1}(\Delta_y(k-\frac{1}{2}) \leq \Pi_{1,1}^{(r]}(t) + \Pi_{1,2}^{(r)}(t) < \Delta_y(k+\frac{1}{2}))}{\sum_{r=1}^{R}\mathbb{1}(\Pi^{(r)}(t) \in \mathring{\mathbb{F}}_3)} \\ \hat{g}_2(k\Delta_y) &= \frac{1}{\Delta_y}\frac{\sum_{r=1}^{R}\mathbb{1}(\Delta_y(k-\frac{1}{2}) \leq \Pi_{1,2}^{(r)}(t) < \Delta_y(k+\frac{1}{2}))}{\sum_{r=1}^{R}\mathbb{1}(\Pi^{(r)}(t) \in \mathring{\mathbb{F}}_3)}\end{aligned}$$

for $k = 1, \ldots, N_y - 1$ (we took $N_y = 50$ in practice). The estimates for $k = 0$ and $k = N_y$ were obtained from the same principle, using the replicates leading to a frequency included in the open intervals $(0, \frac{\Delta_y}{2})$ and $(1 - \frac{\Delta_y}{2}, 1)$ respectively. We computed $\hat{g}_P$ and $\hat{g}_1$ for several values of the population size N and found that $N = 200$ was large enough to get accurate estimates of the exact diffusion densities. The following results were thus obtained for $N = 200$.

We compared the density estimates $\hat{g}_P$ and $\hat{g}_1$ with the approximate densities

$$\tilde{g}_P(p) = \int_0^p dy_1 \int_0^{1-p} \tilde{g}(y_1, p - y_1, y_3)dy_3,\ 0 < p < 1$$

and

$$\tilde{g}_2(y_2) = \int_0^{1-y_2} dy_1 \int_0^{1-y_1-y_2} \tilde{g}(y_1, y_2, y_3)dy_3,\ 0 < y_2 < 1$$

These densities were obtained by numerical integration of the function $\tilde{g}$ returned by our algorithm. For a diffusion process with parameters ρ and τ, we used $c = \frac{\rho}{2N}$ and $t = 2N\tau$ for simulating the Wright-Fisher process.

Hellinger's distances between $\tilde{g}_P$ and $\hat{g}_P$ and between $\tilde{g}_2$ and $\hat{g}_2$ are presented in Tables 3 and 4 respectively. Several values of τ (from 0.01 to 0.5), ρ (from 1 to 10) and y^0 ($(0.15, 0.05, 0.25)$ or $(0.10, 0.01, 0.30)$) were tried. For each configuration of parameter values, the result in Table 3 and the one in Table 4 were computed from the same joint densities $\tilde{g}$ and $\hat{g}$ (in particular the same Monte Carlo replicates were used for both tables). A graphical representation of the several densities we computed is provided in Figure 5.2 in the case of $\tau = 0.1$, $\rho = 10$ and $y^0 = (0.15, 0.05, 0.25)$.

Tables 3 and 4 reveal that our method provides a good approximation of the exact diffusion density. The accuracy was very high for $\tau \leq 0.1$ and $y^0 = (0.15, 0.05, 0.25)$. As already observed and discussed in previous tables, the accuracy was lower when the initial haplotype frequencies were taken closer to the boundary ($y^0 = (0.10, 0.01, 0.30)$) and it was greater with initial frequencies in the middle of $\mathbb{F}_3$ ($y^0 = (0.25, 0.25, 0.25)$, results not shown). Contrary to what was observed in previous tables, the accuracy was here lower for $\tau = 0.5$ than for $\tau = 0.1$. It is not surprising because the two-locus diffusion

with $\rho \neq 0$ has a non zero drift $b(y)$, which was not the case in the one-locus diffusion studied in Table 1 or in the two-locus diffusion with $\rho = 0$ studied in Table 2. The approximate boundary conditions are thus never satisfied here, even for large values of τ. For computational reasons, we could not test the accuracy of $\tilde{g}$ for values of τ greater than 0.5, but it is very likely that the results would be worse than for $\tau = 0.5$. Given these few observations, we could finally expect the accuracy of our approximate density to decrease as ρ increases. It is actually what happens for the frequency of allele 1 at first locus (Table 3). However, the pattern of errors observed for the frequency of haplotype $(1, 2)$ (Table 4) is a little different. One explanation here may be that large values of ρ also have a positive influence on the results because they contribute to create more haplotypes $(1, 2)$ and thus attenuate the weight of the boundary points, as observed in Figure 5.1.

5.5 Example

Here we come back more precisely to the problem of linkage disequilibrium measures that we introduced in Section 5.1, and focus more particularly on the measure $r^2_{i,j}(t)$ given in (5.1). In the case of two bi-allelic loci we actually know that $r^2_{1,1}(t) = r^2_{1,2}(t) = r^2_{2,1}(t) = r^2_{2,2}(t) = r^2(t)$. Sved (1971) provided an approximation of the expected value of this measure as a function of N, c and t:

$$\mathbb{E}[r^2(t)] = \frac{1}{1 + 4Nc\frac{(1-c/2)}{(1-c)^2}}(1 - ((1 - \frac{1}{2N})(1-c)^2)^t)$$

This equation holds under a two-locus Wright-Fisher model without selection or mutation and under the hypothesis that there is no linkage disequilibrium at time 0, i.e. $\mathbb{E}[r^2(0)] = 0$. This formula quantifies the creation of linkage disequilibrium by random drift, and is thus widely used by geneticists.

We compared Sved's formula with the value obtained from our method for $N = 100$, $t = 20$ and several values of c from 0cM to 2cM. We chose an haplotype frequency vector $y^0 = (0.25, 0.25, 0.25)$ and studied the case without selection as well as one case with selection. For this latter case we considered a simple model of directional selection, as described in (Ethier and Nagylaki, 1989). The fitness of allele 1 at first locus was $\sigma = 20$ (in the diffusion model) while other allele fitnesses were equal to 0. At first locus, allele 1 was thus conferred a selective advantage, while the second locus was assumed neutral. Our numerical scheme could be applied exactly as in the neutral case, with only a slight modification of coefficient $\breve{b}(y)$.

As we can see in Figure 5.3, Sved's formula gives slightly larger values of $\mathbb{E}[r^2(t)]$ than

our method with no selection. This is in good agreement with Monte Carlo simulations of Sved (1971) himself, that showed that his formula was an overestimation of $\mathbb{E}[r^2(t)]$ for $N = 50$, $t = 50$ and $c \leq 5$cM. For the parameter values of Figure 5.3, our method is thus more accurate than Sved's formula. Moreover, contrary to Sved's formula, our method can also be applied with weak selection, and the case of $\sigma = 20$ shows that this factor is not negligible.

5.6 Computational complexity

For most values of N, the algorithm we proposed has a much lower computational complexity than direct methods based on the Wright-Fisher model. Indeed the number of iterations needed by our algorithm is $\tau/\Delta_\tau = (t/2N)/\Delta_y^2 = tN_y^2/(2N)$ and each iteration requires a number of operations of order N_y^3 since it computes the value at each discretization point. The complexity is consequently of order tN_y^5/N. In comparison, the direct computation of the Wright-Fisher distribution would require to compute the power t of the transition matrix. Since the size of this matrix is of order N^3, the complexity would be at least of order $(N^3)^2$. As a value of $N_y = 100$ is sufficient to get a good accuracy, our method is clearly faster. Alternatively, we could also estimate the Wright-Fisher distribution by Monte Carlo simulations. To compare it with our method, we would estimate the same number of probabilities, i.e. N_y^3. Given that Monte Carlo schemes converge at speed $1/\sqrt{R}$, where R is the number of replicates, an accuracy of order $\mathcal{O}(\Delta_y)$ would require an average of N_y^2 replicates by point. Then each replicate consists in simulating a population over t generations, that is simulating t multinomial distributions with samples of size N. It requires consequently tN operations, and the total complexity is of order tNN_y^5. The advantage of our method for computing the density is thus obvious, and is increased for large populations. If we are only interested in one specific moment of the density (as in Section 5.5 for instance), the complexity of the Monte Carlo method is reduced to tNN_y^2, while the complexity of our algorithm is the same as before. But as soon as $N \geq 1000$ our method is still more efficient.

Actually for large values of N, simulating the whole population under a Wright-Fisher model cannot be competitive. A reasonable alternative to our method would be to simulate the diffusion process, which requires about $N_y^2 t/(2N)$ operations for one trajectory. For computing one moment of the distribution, N_y^2 trajectories are generally enough and this approach could thus be faster than ours. For evaluating the whole distribution our method is still more efficient. Simulating the haplotype frequencies along the ancestry graph of a given sample at time t is an other way to reduce the complexity of the problem

(Gasbarra et al., 2005). However, the result obtained by such methods cannot exactly be compared to the result that we get, which is at population level.

All these complexity results are summarized in Table 4 and evaluated for two sets of parameter values. Practically, the computing time needed for our method to obtain the approximate joint density with $\tau = 0.1$ and $N_y = 100$ was about 45 minutes on a Intel XEON processor $4 \times 2.4GHz$.

5.7 Discussion

In this study we were interested in the probability distribution of haplotype frequencies under the two-locus diffusion model with bi-allelic loci. We presented a method aiming at approximating the conditional joint density of these frequencies given that no haplotype has been lost. This method is based on a numerical solution of the Kolmogorov forward equation. It can be applied to evaluate transient densities, and can deal with models including selection or mutation. Several comparisons and tests were provided, that evaluate the accuracy of this approximate method.

Results of Sections 5.4 clearly indicate that our numerical solution of the Kolmogorov forward equation introduces no significant error for the conditional density of $Y(\tau)$ if we consider a model without selection or mutation, a diffusion time $\tau \leq 0.1$ and if our initial haplotype frequency vector y^0 is not too close to the boundaries of $\mathbb{F}_3$. For initial haplotype frequencies very close to the boundaries of $\mathbb{F}_3$ or for larger evolution times, this approximation remains reasonable but should be considered with more caution: in that case Hellinger's distances between the exact and approximate solutions can reach an order of 10^{-2}, which may be problematic in some applications. The influence of selection or of mutations between alleles on the accuracy of our method was not tested, essentially because of the difficulty to find an exact result that we could use for the comparison. Future studies based on intensive simulations could help to answer this question. However, it should be noted that introducing mutation rates between alleles would make the fixation probabilities vanish. The problem caused by the behavior at the boundaries would then be reduced, and the accuracy would certainly increase.

In light of these considerations about the accuracy of the approximation, and of the general assumptions made by the model, which are the situations where our method can be used? First of all, the description of a population by a vector of haplotype frequencies of fixed small size implies that no new allele can be created between times 0 and τ. Besides, the diffusion approximation also has some consequences on the values that can be used for the genetical parameters. Theorem 2 actually implies that the limit recombination

rate considered is around 1, which in practice corresponds to closely linked loci. The same remarks can be done concerning the selection and mutation rates, as pointed out by Ethier and Norman (1977). For example, the most recent generations in many animal breeds used in agriculture don't satisfy this hypothesis because the selection pressure they are subject to is too strong. Finally, our numerical solution is accurate provided τ is at most of order 0.1, which actually means t of order $0.1N$. This leads us to consider relatively short population histories, typically shorter than the ones used by coalescent methods.

When the above conditions are reached, our method is a powerful tool for studying the evolution of haplotype frequencies. It is particularly convenient for problems where haplotype frequencies are still in transient state. The forward version, on which we focused in this paper, is a competitive alternative to traditional forward simulation programs, as underlined in section 5.6. The potential applications are numerous (see (Gasbarra et al., 2005) for a short review). The backward version also opens many perspectives, and will motive further papers.

Our method can be particularly useful, for instance, in the field of linkage disequilibrium gene mapping. Indeed, the small values of τ we considered correspond to the ones usually found in fine mapping studies of disease genes. In several of these studies (Kaplan et al., 1995; Graham and Thompson, 1998; Morris et al., 2000), the population effective size was estimated around 10000 and the number of generations since causal mutation around 200, which gives $\tau = 0.01$. In this case the values we chose for ρ give $c = 0.005$ centi Morgan (cM), 0.025cM and 0.05cM. It corresponds to the order of distances considered in reality. One first example of result on the property of linkage disequilibrium measures was provided in Section 5.5.

The major drawback of our method is the approximation we have to do about the boundary conditions, which for some parameter values introduces significant errors. However, it is to be noticed that these errors mainly occur in the case of parameter values for which the diffusion limit itself should not be used to approximate the Wright-Fisher process: extreme initial and final allele frequencies (Bürger and Ewens, 1995) and large values of ρ. Moreover, these errors are reduced in the backward problem, where the boundary values are generally more fully determined. Last but not least, we observed that despite the approximate boundary conditions our solution was convergent as τ tends to infinity in the one-locus model or the two-locus model with $\rho = 0$. These results are very encouraging and let us think that we could find boundary conditions which also provide better results for $\rho \neq 0$. This topic will motive further investigations.

Extending our method to models with more than two loci would also be useful in the present genetical context where more and more genomic information is available. In

theory this extension would be possible, but from a numerical point of view it seems more difficult. For solving evolution equations in dimension greater than 3, simulating the diffusion process should be more efficient.

Table 1: Numerical solution versus exact solution for one locus

	τ				
y^0	0.01	0.1	0.5	1	2
0.5	9.90e-03	5.91e-05	5.66e-07	1.32e-09	2.09e-12
0.2	1.14e-02	6.82e-05	2.41e-05	3.29e-06	6.06e-08
0.05	3.20e-03	9.10e-04	5.34e-05	7.73e-06	1.37e-07
0.01	7.00e-03	2.30e-03	6.32e-05	9.26e-06	1.62e-07

Hellinger's distance between the exact density g of the frequency of allele 1 under a one-locus diffusion model and the approximation $\tilde{g}$ of this density computed by our method for different values of the time τ and of the initial allele frequency y^0. $\Delta_y = 1/100$ and $\Delta_\tau = \Delta_y^2$ for the numerical scheme

Table 2: Numerical solution versus exact solution for two loci with $\rho = 0$

	τ		
y^0	0.1	0.5	1
$(0.15, 0.05, 0.25)$	9.41e-04	1.28e-04	2.73e-06
$(0.10, 0.01, 0.30)$	6.80e-02	1.70e-03	3.20e-05

Hellinger's distance between the exact joint density g of haplotype frequencies under a two-locus diffusion model with recombination rate $\rho = 0$ and the approximation $\tilde{g}$ of this density computed by our method for different values of the time τ and of the initial allele frequency y^0. $\Delta_y = 1/100$ and $\Delta_\tau = \Delta_y^2$ for the numerical scheme.

Table 3: Marginal allele density

y^0	$(0.15, 0.05, 0.25)$			$(0.10, 0.01, 0.30)$		
	τ			τ		
ρ	0.01	0.1	0.5	0.01	0.1	0.5
1	4.54e-04	6.90e-03	1.41e-02	6.60e-03	1.82e-02	2.63e-02
5	1.00e-03	6.90e-03	1.66e-02	8.40e-03	2.23e-02	3.07e-02
10	9.00e-04	7.70e-03	7.12e-02	1.10e-02	2.64e-02	4.87e-02

Hellinger's distance between the density $\hat{g}_P$ of the frequency of allele 1 at first locus under a two-locus Wright-Fisher model with population size $N = 200$ and the corresponding density $\tilde{g}_P$ under a two-locus diffusion model. $\hat{g}_P$ is estimated by Monte Carlo using 10^5 replicates and $\tilde{g}_P$ is computed by our numerical scheme with $\Delta_y = 1/100$ and $\Delta_\tau = \Delta_y^2$. Several values of the time τ, of the recombination rate ρ and of the initial haplotype frequencies y^0 were used.

Table 4: Marginal haplotype density

y^0	$(0.15, 0.05, 0.25)$			$(0.10, 0.01, 0.30)$		
	τ			τ		
ρ	0.01	0.1	0.5	0.01	0.1	0.5
1	1.51e-02	5.10e-03	1.28e-01	4.21e-02	8.10e-02	2.32e-01
5	9.30e-03	7.36e-04	5.50e-03	7.20e-02	1.40e-02	1.36e-02
10	6.40e-03	1.60e-03	2.41e-02	1.37e-01	6.60e-03	2.24e-02

Hellinger's distance between the density $\hat{g}_2$ of the frequency of haplotype $(1, 2)$ under a two-locus Wright-Fisher model with population size $N = 200$ and the corresponding density $\tilde{g}_2$ under a two-locus diffusion model. $\hat{g}_2$ is estimated by Monte Carlo using 10^5 replicates and $\tilde{g}_2$ is computed by our numerical scheme with $\Delta_y = 1/100$ and $\Delta_\tau = \Delta_y^2$. Several values of the time τ, of the recombination rate ρ and of the initial haplotype frequencies y^0 were used.

Table 5: Computational complexity

Population parameters	Method			
	FD	TM	MCWF	MCD
Whole density				
General case	tN_y^5/N	N^6	tNN_y^5	tN_y^7/N
$N = 1000$, $t = 100$, $N_y = 100$	10^9	10^{18}	10^{15}	10^{13}
$N = 10000$, $t = 100$, $N_y = 100$	10^8	10^{24}	10^{12}	10^{12}
One specific moment				
General case	tN_y^5/N	N^6	tNN_y^2	tN_y^4/N
$N = 1000$, $t = 100$, $N_y = 100$	10^9	10^{18}	10^9	10^7
$N = 10000$, $t = 100$, $N_y = 100$	10^8	10^{24}	10^{10}	10^6

Computational complexity for different methods computing the distribution of haplotype frequencies under the Wright-Fisher model. FD : our method (based on diffusion approximation and finite differences). TM : direct method based on the transition matrix of the Wright-Fisher process. MCWF : Monte Carlo simulations of the Wright-Fisher process. MCD : Monte Carlo simulations of the diffusion process.

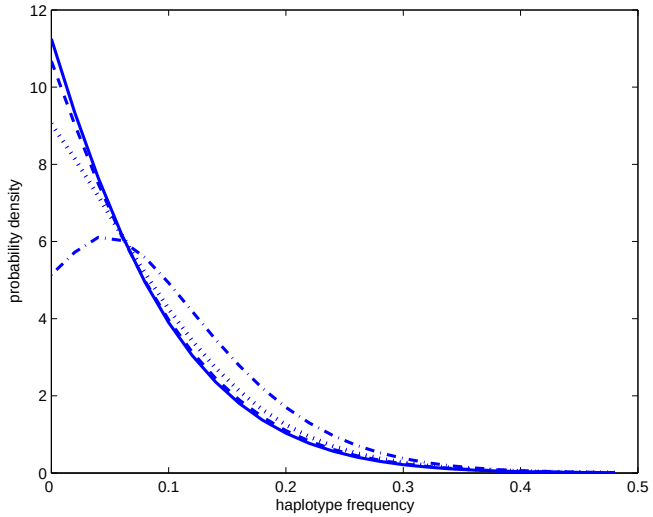
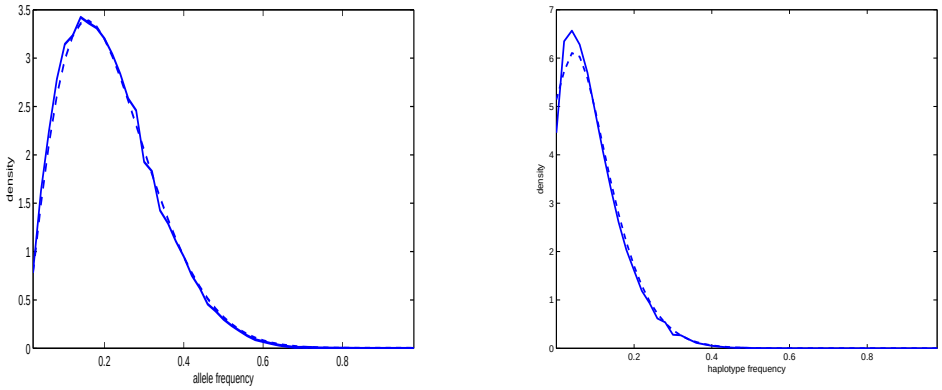


Figure 5.1: Exact marginal density g_2 of the frequency of haplotype $(1, 2)$ for the two-locus diffusion model with recombination rate $\rho = 0$ (solid line) and approximation $\tilde{g}_2$ of this density obtained by our method with $\rho = 0$ (dash line), 5 (dotted line) and 10 (dash-dotted line). The time is $\tau = 0.1$ and the initial haplotype frequencies are $y^0 = (0.15, 0.05, 0.25)$.



(a) $\tilde{g}_P$ and $\hat{g}_P$ (b) $\tilde{g}_2$ and $\hat{g}_2$

Figure 5.2: Comparison between the density of the two-locus diffusion model with time $\tau = 0.1$ and recombination rate $\rho = 10$ (dash line) and the density of the two-locus Wright-Fisher model with population size $N = 200$, time $t = 40$ and recombination rate $c = 0.025M$ (solid line). The diffusion density is approximated by our method with $\Delta_y = 1/100$ and $\Delta_\tau = \Delta_y^2$. The Wright-Fisher density is estimated by Monte Carlo with $N_y = 50$ and 10^5 iterations. The initial haplotype frequencies are $y^0 = (0.15, 0.05, 0.25)$. (a) marginal densities $\tilde{g}_P$ and $\hat{g}_P$ of allele 1 at first locus. (b) marginal densities $\tilde{g}_2$ and $\hat{g}_2$ of haplotype $(1, 2)$.

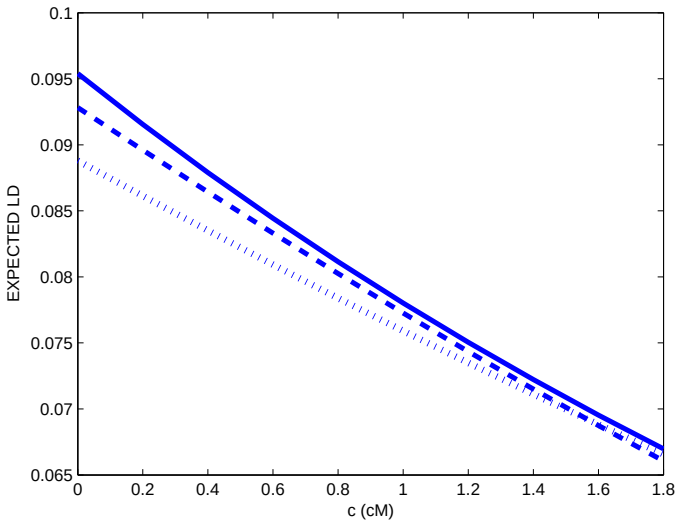


Figure 5.3: $\mathbb{E}[r^2_{i,j}(t)]$ under the two-locus Wright-Fisher model as a function of the recombination rate c for time $t = 20$ and population size $N = 100$. Solid line : Sved's formula. Dash line : our method with selection rate $\sigma = 0$. Dotted line : our method with selection rate $\sigma = 20$.

Numerical solution of the Kolmogorov forward equation for one locus

Here is the MATLAB program we used in Section 5.4.1 to compute the approximate diffusion density. Except for accuracy tests as we did, it is not very useful because the exact solution is known. Nevertheless it gives a good idea about how our algorithm for two loci works, and it is much shorter and easier to understand. The main program is *densite.m* and it appeals the subprogram *op.m*.

```
function uT = densite(u0,T,Nt,I)
% this program computes the probability density of allele
% frequency at time T under a one-locus diffusion model
% without selection or mutation. It resolves the
% Kolmogorov forward equation by a finite difference scheme
% Boundary conditions : second order derivative vanishes
% uT : density at time T
% u0 : density at time 0
% T : final time
% Nt : number of discretization intervals for [0,T]
% I : number of discretization intervals for [0,1]
% vector index :
% index 1 corresponds to y=0
% index I+1 corresponds to y=1

h=1/I; % h is the discretization step for [0,1]
dT=T/Nt; % dT is the discretization step for [0,T]
u_av=u0;
for t=1:Nt
  % at each iteration, u_ap is computed from u_av
  u_ap=zeros(I+1,1);
  % for interior points we apply formula (6)
  for i=2:I
        u_ap(i)=u_av(i)+dT*op(u_av,i,h);
  end
  % for boundary points we vanish the second order derivate
  % y=0
```

```
  u_ap(1)=2*u_ap(2)-u_ap(3);
  % y=1
  u_ap(I+1)=2*u_ap(I)-u_ap(I-1);
  % for the next iteration we restart from the density
  % we have just computed
  u_av=u_ap;
end
uT=max(u_ap,0); % just to avoid artefact negative values



function y = op(u,i,h)
% computes Lu((i-1)h) where L is the discrete version
% of the KFE operator
% u : a real function
% i : index vector of the point where we apply the opeator
% h : discretization step for [0,1]

x=(i-1)*h; % computation of the point corresponding to index i
% the first term of the operator is a function of u
y=-u(i);
% the second term of the operator is a function of
% the first derivative of u
b=1-2*x;
if b>0
y=y+b*(u(i+1)-u(i))/h;
% (u(i+1)-u(i))/h is an approximation of
% the first derivative of u at (i-1)*h
else
y=y+b*(u(i)-u(i-1))/h;
end
% the third term of the operator is a function of
% the second derivative of u
y=y+(x*(1-x)*(u(i+1)-2*u(i)+u(i-1)))/(2*h*h);
% (u(i+1)-2*u(i)+u(i-1))/(2*h*h) is an approximation of
% the second  derivative of u at (i-1)*h
```

References

Barbour, A., Ethier, S., and Griffiths, R. (2000). A transition density expansion for a diffusion model with selection. *Ann. Appl. Probab.*, *10*, 123-162.

Billingsley. (1968). *Convergence of probability measures.* John Wiley and Sons, Inc.

Bürger, R., and Ewens, W. (1995). Fixation probabilities of additive alleles in diploid populations. *J. Math. Biol.*, *33*, 557-575.

Cierco-Ayrolles, C., Abdallah, J., Boitard, S., Chikhi, L., Rochambeau, H. de, Tsitrone, A., et al. (2004). On linkage disequilibrium measures: Methods and applications. In (Vol. 1, p. 151-180). Research Signpost, India.

Crow, J., and Kimura, M. (1970). *An introduction to population genetics theory.* Harper and Row, New York.

Devlin, B., and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, *29*, 311-322.

Erdélyi, A. (1953). *Higher transcendental functions* (Vol. 2). McGraw-Hill, New York.

Ethier, S., and Kurtz, T. (1986). *Markov processes. characterization and convergence.* John Wiley and Sons, Inc.

Ethier, S., and Nagylaki, T. (1980). Diffusion approximations of markov chains with two time scales and applications to population genetics. *Adv. Appl. Probab.*, *12*, 14-49.

Ethier, S., and Nagylaki, T. (1989). Diffusion approximations of the two-locus wright-fisher model. *J. Math. Biol.*, *27*, 17-28.

Ethier, S., and Norman, M. (1977). An error estimate of the diffusion approximation of the diffusion process. *Proc. Natl. Acad. Sci. USA*, *74*, 5096-5098.

Fisher, R. (1922). On the dominance ratio. In *Proceedings of the royal society of edinburgh* (Vol. 42, p. 321-341).

Gasbarra, D., Sillanpää, M., and Arjas, E. (2005). Backward simulation of ancestors of sampled individuals. *Theor. Popul. Biol.*, *67*, 75-83.

Graham, J., and Thompson, E. (1998). Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am. J. Hum. Genet.*, *63*, 1517-1530.

Griffiths, R. (1979). A transition density expansion for a multi-allele diffusion model. *Adv. Appl. Probab.*, *11*, 310-325.

Griffiths, R. (1981). Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.*, *19*, 169-186.

Hudson, R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, *109*, 611-631.

Kaplan, N., Hill, W., and Weir, B. (1995). Likehood methods for locating disease genes

in nonequilibrium populations. *Am. J. Hum. Genet.*, *56*, 18-32.

Karlin, S., and McGregor, J. (1968). Rates and probabilities of fixation for two locus random mating finite populations without selection. *Genetics*, *58*, 141-159.

Karlin, S., and Taylor, H. (1981). *A second course in stochastic processes.* Academic Press, Inc.

Kimura, M. (1955a). Random genetic drift in a tri-allelic locus; exact solution with a continuous model. *Biometrics*, *12*, 57-66.

Kimura, M. (1955b). Solution of a process of random genetic drift with a continuous model. *Proc. Nat. Acad. Sci. USA*, *41*, 144-150.

Kimura, M. (1955c). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology*, *20*, 33-53.

Kimura, M. (1964). Diffusion models in population genetics. *J. Appl. Probab.*, *1*, 177-232.

Litler, R. (1973). Linkage disequilibrium in two-locus, finite, random mating models without selection or mutation. *Theor. Popul. Biol.*, *4*, 259-275.

Littler, R., and Fackerell, E. (1975). Transition densities for neutral multi-allele diffusion models. *Biometrics*, *31*, 117-123.

Lucquin, B., and Pironneau, O. (1995). *Introduction au calcul scientifique.* Masson.

Mano, S. (2005, December). Random genetic drift and gamete frequency. *Genetics*, *171*, 2043-2050.

Morris, A., Whittaker, J., and Balding, D. (2000). Bayesian fine-scale mapping of disease loci by hidden markov models. *Am. J. Hum. Genet.*, *67*, 155-169.

Nordborg, M. (2001). Coalescent theory. In D. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of statistical genetics* (p. 179-212). Wiley.

Ohta, T., and Kimura, M. (1969). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, *63*, 229-238.

Stroock, D., and Varadhan, S. (1997). *Multidimensionnal diffusion processes.* Springer-Verlag.

Sved, J. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.*, *2*, 125-141.

Tier, C., and Keller, J. B. (1978). Asymptotic analysis of diffusion equations in population genetics. *SIAM J. Appl. Math.*, *34* (3), 549-576.

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

Chapitre 6

Calcul de la vraisemblance par l'équation rétrograde de Kolmogorov

Au cours du chapitre 4, nous avons vu qu'un moyen d'estimer la position d'un QTL était de calculer la vraisemblance de chaque position x en intégrant par rapport à la distribution des fréquences d'haplotypes $\Pi(t)$ dans la population, ce qui donne

$$\mathcal{L}(x \mid \mathcal{D}) = \mathbb{E}[\mathcal{L}(x \mid \mathcal{D}, \Pi(t))]$$

La méthode numérique présentée au chapitre 5 nous fournit désormais un moyen de calculer la densité de $\Pi(t)$, et on peut intégrer numériquement par rapport à cette densité pour calculer $\mathcal{L}(x \mid \mathcal{D})$ dans le cas d'un modèle à deux loci. En réalité, on peut même calculer directement $\mathcal{L}(x \mid \mathcal{D})$ en passant par l'équation rétrograde de Kolmogorov. L'objet de ce chapitre est de comparer les courbes de vraisemblance obtenues par cette méthode à celles obtenues par d'autres approximations classiques.

6.1 Modèle

Le problème considéré ici est l'estimation du taux de recombinaison c entre un gène d'intérêt biallélique d'allèles Q et q et un marqueur biallélique d'allèles 1 et 2. Puisque la position du marqueur est connue, estimer c revient à estimer la position x du gène d'intérêt. Pour simplifier, on supposera que la relation entre génotypes et phénotypes est déterministe, de sorte qu'on observe directement les effectifs des haplotypes $(Q,1)$, $(Q,2)$, $(q,1)$ et $(q,2)$ dans l'échantillon. Ces effectifs sont notés $n_{Q,1}, \ldots, n_{q,2}$ et l'ensemble des données disponibles $\mathcal{D}$ est simplement constitué du vecteur $n = (n_{Q,1}, \ldots, n_{q,2})$. Le vecteur des fréquences d'haplotypes dans la population à la génération t est $\Pi(t) = (\Pi_{Q,1}(t), \ldots, \Pi_{q,2}(t))$.

Il suit le modèle de Wright-Fisher à deux loci décrit en 2.1.2.

On suppose que les haplotypes observés dans l'échantillon sont issus d'un tirage aléatoire avec remise parmi les haplotypes présents dans la population à la génération t, si bien que

$$\begin{aligned}\mathbb{P}(n \mid \Pi(t)) &= C_n \Pi_{Q,1}(t)^{n_{Q,1}} \Pi_{Q,2}(t)^{n_{Q,2}} \Pi_{q,1}(t)^{n_{q,1}} \Pi_{q,2}(t)^{n_{q,2}} \\ &= C_n \prod_{i_1=Q,q} \prod_{i_2=1,2} \Pi_{i_1,i_2}(t)^{n_{i_1,i_2}}\end{aligned}$$

où $C_n = \frac{(2N_s)!}{n_{Q,1}!n_{Q,2}!n_{q,1}!n_{q,2}!}$ et $2N_s = n_{Q,1} + n_{Q,2} + n_{q,1} + n_{q,2}$. N_s est la taille de l'échantillon, qui compte $2N_s$ haplotypes. Or $\Pi(t)$ est une variable cachée dont la loi dépend de c, t, de la taille de la population $2N$ et du vecteur des fréquences d'haplotypes initiales π^0. La vraisemblance de ces paramètres peut donc s'écrire

$$\mathcal{L}(c, t, N, \pi^0 \mid n) = \mathbb{E}[\prod_{i_1=Q,q} \prod_{i_2=1,2} \Pi_{i_1,i_2}(t)^{n_{i_1,i_2}}] \tag{6.1}$$

On suppose en fait que t et N sont des constantes connues. D'autre part, conformément aux hypothèses classiques de la cartographie par déséquilibre de liaison (voir section 4.3), on considère que l'instant 0 correspond à un événement démographique ou génétique créant un fort déséquilibre de liaison entre les 2 loci. En particulier, l'allèle le plus récent du gène d'intérêt, disons Q, n'est alors associé qu'à un seul des deux allèles du marqueur. On a donc $\pi^0_{Q,1} = 0$ ou $\pi^0_{Q,2} = 0$

6.2 Estimateurs

Dans le cadre du modèle ci-dessus, on souhaite estimer c et π^0 par les estimateurs de maximum de vraisemblance $\hat{c}$ et $\hat{\pi}^0$ tels que

$$\mathcal{L}(\hat{c}, t, N, \hat{\pi}^0 \mid n) = \sup_{\pi^0_{Q,1}=0 \ \cup \ \pi^0_{Q,2}=0} \mathcal{L}(c, t, N, \pi^0 \mid n)$$

Nous comparons trois estimateurs, qui diffèrent par la façon de calculer $\mathcal{L}(c, t, N, \pi^0 \mid \mathcal{D})$.

6.2.1 Méthode de Monte Carlo

Une première possibilité consiste à estimer la vraisemblance par Monte Carlo. Pour des valeurs de N, t, c et π^0 fixées, on simule des trajectoires de fréquences $\Pi_m(.)$, $m = 1, \ldots, M$ sous un modèle de Wright-Fisher à deux loci. A chaque génération t,

on peut tirer le vecteur de fréquences $\Pi_m(t+1)$ à l'aide de l'algorithme décrit en 2.1.2. On peut aussi utiliser un simulateur de loi multinomiale déjà programmé puisque pour le modèle à deux loci on a une expression explicite des probabilités de transition donnée par l'équation (2.1). On en déduit

$$\mathcal{L}^{(MC)}(c,t,N,\pi^0 \mid n) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}(c,t,N,\pi^0 \mid n, \Pi_m(t)) \tag{6.2}$$

A priori, il faudrait effectuer ce calcul pour un grand nombre de valeurs de c et π^0 afin de chercher les valeurs qui maximisent $\mathcal{L}^{(MC)}$. Cela serait extrêmement long, et on se contente donc en pratique de calculer cette vraisemblance pour $\pi^0 = (0, \hat{\pi}^0_{Q,.}, \hat{\pi}^0_{.,1}, 1-\hat{\pi}^0_{Q,.}-\hat{\pi}^0_{.,1})$ et $\pi^0 = (\hat{\pi}^0_{Q,.}, 0, \hat{\pi}^0_{.,1}-\hat{\pi}^0_{Q,.}, 1-\hat{\pi}^0_{.,1})$, où $\hat{\pi}^0_{Q,.}$ et $\hat{\pi}^0_{.,1}$ sont les estimateurs de maximum de vraisemblance des fréquences alléliques initiales $\pi^0_{Q,.} = \pi^0_{Q,1}+\pi^0_{Q,2}$ et $\pi^0_{.,1} = \pi^0_{Q,1}+\pi^0_{q,1}$. D'après l'étude du modèle de Wright-Fisher à un locus, il est clair que $\hat{\pi}^0_{Q,.} = \frac{n_{Q,1}+n_{Q,2}}{2N_s}$ puisque $n_Q = n_{Q,1}+n_{Q,2}$ est issu d'un tirage binomial $\mathcal{B}(2N_s, \Pi_{Q,.}(t))$ et que d'autre part $\mathbb{E}[\Pi_{Q,.}(t)] = \pi^0_{Q,.}$. On obtient de même que $\hat{\pi}^0_{.,1} = \frac{n_{Q,1}+n_{q,1}}{2N_s}$. Sachant que la somme des 4 fréquences d'haplotypes doit être égale à 1, les deux valeurs de π^0 proposées ci-dessus sont les seules permettant de respecter les estimateurs des fréquences alléliques et les conditions respectives $\pi^0_{Q,1} = 0$ et $\pi^0_{Q,2} = 0$.

6.2.2 Approximation du premier ordre

Une deuxième méthode, proposée par Xiong et Guo (1997) dans le cadre d'études cas contrôle pour localiser des gènes de maladies rares, est de calculer une approximation au premier ordre de la vraisemblance définie par l'équation (6.1). On obtient ainsi

$$\mathcal{L}^{(1)}(c,t,N,\pi^0 \mid n) = \prod_{i_1=Q,q}\prod_{i_2=1,2} \mathbb{E}[\Pi_{i_1,i_2}(t)^{n_{i_1,i_2}}] \tag{6.3}$$

Dans le cadre du modèle de Wright-Fisher, on sait calculer de manière exacte les espérances des fréquences d'haplotypes. On trouve

$$\mathbb{E}[\Pi_{i_1,i_2}(t)] = \pi^0_{i_1,i_2} - cd^0_{i_1,i_2}\frac{1-\theta^{t+1}}{1-\theta} \tag{6.4}$$

où $d^0_{i_1,i_2} = \pi^0_{i_1,i_2} - \pi^0_{i_1,.}\pi^0_{.,i_2}$ et $\theta = (1-c)(1-\frac{1}{2N})$. Ici le calcul de $\mathcal{L}^{(1)}(c,t,N,\pi^0 \mid n)$ est très rapide donc on peut se permettre d'essayer de nombreuses valeurs de π^0 telles que $\pi^0_{Q,1} = 0$ ou $\pi^0_{Q,2} = 0$.

Pour une valeur de π^0 fixée, on constate qu'estimer c en maximisant $\mathcal{L}^{(1)}(c,t,N,\pi^0 \mid n)$

revient à prendre l'estimateur empirique solution de l'équation

$$(1-c)^t(1-\frac{1}{2N})^t = \frac{\hat{D}(t)}{d^0}$$

où $\hat{D}(t) = \frac{n_{Q,1}}{N_s} - \frac{n_{Q,1}+n_{Q,2}}{N_s}\frac{n_{Q,1}+n_{q,1}}{N_s}$ est l'estimateur empirique du déséquilibre de liaison au temps t. Rappelons que pour le modèle à deux loci, $D_{i_1,i_2}(t)$ ne dépend du couple d'allèles (i_1,i_2) que par son signe. Le quotient $\frac{D_{i_1,i_2}(t)}{d^0_{i_1,i_2}}$ ne dépend donc pas du couple (i_1,i_2), d'où la notation simplifiée $\frac{D(t)}{d^0}$.

6.2.3 Équation rétrograde de Kolmogorov

Dans la dernière méthode, la loi de $\Pi(t)$ associée à un taux de recombinaison c est approchée par la loi du processus de diffusion $Y(\tau)$ décrit en 5.2.2, pour $\tau = \frac{t}{2N}$ et un taux de recombinaison $\rho = 2Nc$. On utilise donc l'approximation

$$\mathcal{L}(c,t,N,\pi^0 \mid n) \approx \mathcal{L}(\rho,\tau,\pi^0 \mid n) = \mathbb{E}[\prod_{i_1=Q,q}\prod_{i_2=1,2} Y_{i_1,i_2}(\tau)^{n_{i_1,i_2}}] \tag{6.5}$$

Or $\mathcal{L}(\rho,\tau,\pi^0 \mid n)$ vérifie l'équation rétrograde de Kolmogorov

$$\frac{\partial \mathcal{L}}{\partial \tau} = L\mathcal{L} = \sum_{j=1}^{3} b_j(\pi^0)\frac{\partial \mathcal{L}}{\partial y_j}(\pi^0) + \frac{1}{2}\sum_{j_1,j_2=1}^{3} a^2_{j_1,j_2}(\pi^0)\frac{\partial^2 \mathcal{L}}{\partial y_{j_1}\partial y_{j_2}}(\pi^0)$$

où L, b et a^2 sont le générateur et les coefficients infinitésimaux définis en 5.3 et où, comme dans le chapitre précédent, on note

$$Y(\tau) = (Y_{Q,1}(\tau), Y_{Q,2}(\tau), Y_{q,1}(\tau)) = (Y_1(\tau), Y_2(\tau), Y_3(\tau))$$

Pour des valeurs de ρ et τ fixées, on peut utiliser l'algorithme présenté au chapitre précédent pour calculer $\mathcal{L}(\rho,\tau,\pi^0 \mid n)$ pour tout $\pi^0 \in \mathbb{F}_3$. Il y a quelques modifications simples à opérer sur l'algorithme. Tout d'abord le générateur discrétisé $\tilde{L}^*$ est remplacé par un nouveau générateur $\tilde{L}$ pour tenir compte des différences entre les coefficients des deux équations. D'autre part, la condition initiale devient

$$\mathcal{L}(\rho,0,\pi^0 \mid n) = \prod_{i_1=Q,q}\prod_{i_2=1,2} (\pi^0_{i_1,i_2})^{n_{i_1,i_2}},\ \pi^0 \in \mathbb{F}_3$$

Enfin, si toutes les coordonnées de n sont non nulles, on a les conditions aux bords

$$\mathcal{L}(\rho, \tau, \pi^0 \mid n) = 0 \text{ pour } \pi^0_{Q,\cdot} = 0, \pi^0_{Q,\cdot} = 1, \pi^0_{\cdot,1} = 0 \text{ ou } \pi^0_{\cdot,1} = 1$$

Cela traduit le fait que si à un moment donné un des allèles disparaît, il ne peut pas être recréé et il est alors impossible d'observer tous les types d'haplotypes dans l'échantillon final. Les autres conditions aux bords -celles correspondant aux bords non absorbants en fait- restent les mêmes qu'au chapitre précédent.

On appelle $\mathcal{L}^{(KBE)}(\rho, \tau, \pi^0 \mid n)$ la vraisemblance retournée par cet algorithme. Comme nous obtenons la solution pour toutes les valeurs de π^0, il est facile de prendre ensuite le maximum sur $\pi^0_{Q,1} = 0$ ou $\pi^0_{Q,2} = 0$.

Le tableau 6.1 récapitule les 3 approximations présentées dans cette section pour le calcul de la vraisemblance (6.1).

Approximation	Principe	Référence
$\mathcal{L}^{(MC)}$	Processus de Wright-Fisher Approximation de Monte Carlo	formule (6.2)
$\mathcal{L}^{(1)}$	Processus de Wright-Fisher Equivalent de la vraisemblance quand $N \to \infty$	formules (6.3) et (6.4)
$\mathcal{L}^{(KBE)}$	Processus de diffusion Résolution numérique de l'équation de Kolmogorov	section 6.2.3

TAB. 6.1 – Récapitulatif des 3 approximations utilisées dans ce chapitre pour calculer la vraisemblance $\mathcal{L}(c, t, N, \pi^0 \mid n)$ donnée par la formule (6.1).

6.3 Allure des vraisemblances

Pour $N = 10000$, $t = 100$, $\pi^0 = (0.21, 0, 0.37)$, et un échantillon $n = (29, 3, 58, 59)$ (de taille totale $N_s = 75$), la figure 6.1 montre la courbe de vraisemblance pour c obtenue par chacune des 3 méthodes. Nous avons volontairement pris un échantillon parmi les plus probables, c'est à dire proche de $2N_s\mathbb{E}[\Pi(t)]$. Les valeurs de c ont également été sélectionnées de façon à englober le maximum $\hat{c}$. Enfin, pour éviter de manipuler des valeurs de vraisemblance trop faibles qui pourraient être à l'origine d'erreurs numériques, la vraisemblance calculée n'est pas exactement celle de la formule (6.1) mais plutôt

$$\mathcal{L}'(c, t, N, \pi^0 \mid n) = \mathbb{E}[\prod_{i_1=Q,q} \prod_{i_2=1,2} (\frac{\Pi_{i_1,i_2}(t)}{p_{i_1,i_2}})^{n_{i_1,i_2}}]$$

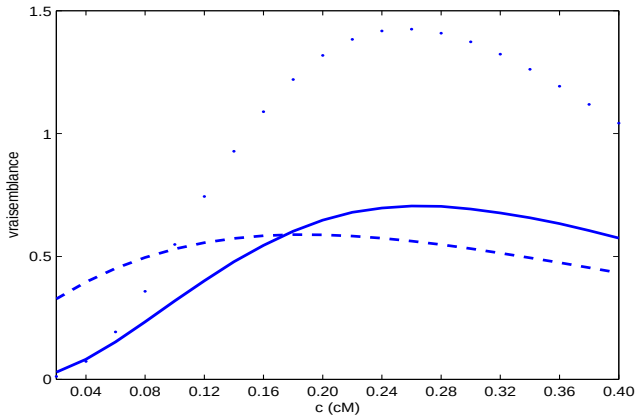


FIG. 6.1 – Courbes de vraisemblance $\mathcal{L}^{(MC)}(c, t, N, \pi^0 \mid n)$ (ligne continue), $\mathcal{L}^{(1)}(c, t, N, \pi^0 \mid n)$ (points), et $\mathcal{L}^{(KBE)}(\rho, \tau, \pi^0 \mid n)$ (ligne en pointillés) en fonction de c. On a $N = 10000$, $t = 100$, $\pi^0 = (0.21, 0, 0.37)$ et $n = (29, 3, 58, 59)$. Pour $\mathcal{L}^{(MC)}(c, t, N, \pi^0 \mid n)$ on simule $M = 300$ populations pour chaque valeur de c.

où $p = 2N_s\mathbb{E}[\Pi(t)] = (0.18, 0.02, 0.37, 0.43)$. Cela revient à calculer la densité de probabilités du vecteur n par rapport à la mesure image de la loi multinomiale $\mathcal{M}(2N_s, p)$ plutôt que par rapport à la mesure de comptage sur $\mathbb{N}^4$. $\mathcal{L}'$ est proportionnelle à $\mathcal{L}$ donc cela ne change rien pour l'estimation.

A condition de simuler suffisamment de fois l'histoire de la population, la vraisemblance obtenue par Monte Carlo peut être considérée comme la courbe de référence. C'est le cas ici, et nous avons vérifié qu'il y avait bien convergence de la méthode. Mais le temps de calcul nécessaire est très important, et on préférerait pouvoir utiliser l'une des deux autres méthodes.

Sur cet exemple, l'approximation d'ordre 1 donne le même maximum que la méthode de Monte Carlo, mais le pic de vraisemblance obtenu est beaucoup plus fort. Cela semble se vérifier pour les quelques autres exemples d'échantillons étudiés. Ceci provient du fait que la convergence de $\mathcal{L}$ vers $\mathcal{L}^{(1)}$ quand N tend vers l'infini est très lente. Même pour $N = 10000$, il y a encore une différence importante entre les deux expressions. Pour les mêmes valeurs de π^0, t, n et le taux de recombinaison $c = 0.26cM$ qui maximise $\mathcal{L}^{(1)}$, la table 6.2 montre qu'il faudrait une population de taille $N = 500000$ pour avoir une bonne adéquation entre $\mathcal{L}^{(1)}$ et $\mathcal{L}^{(MC)}$. En pratique, l'utilisation de l'approximation d'ordre 1 conduirait sans doute à des intervalles de confiance trop étroits, car elle néglige d'une certaine manière la variabilité de l'histoire évolutive.

L'approximation par la diffusion rend mieux compte de cette variabilité évolutive. Cela se retrouve sur la courbe de vraisemblance, dont l'incurvation est plus proche de celle de

N	$\mathcal{L}^{(MC)}(c, t, N, \pi^0 \mid n)$
10 000	0.71
20 000	0.94
50 000	1.21
100 000	1.29
200 000	1.36
500 000	1.40

TAB. 6.2 – Évolution de $\mathcal{L}^{(MC)}(c, t, N, \pi^0 \mid n)$ en fonction de N, pour $t = 100$, $c = 0.26cM$, $\pi^0 = (0.21, 0, 0.37)$ et $n = (29, 3, 58, 59)$. Chaque valeur du tableau est calculée à partir de 200 populations simulées. Pour les paramètres de ce tableau, on a toujours $\mathcal{L}^{(1)}(c, t, N, \pi^0 \mid n) = 1.42$

la courbe obtenue par Monte Carlo. En revanche, le maximum est atteint pour une valeur de c légèrement inférieure. Cette différence ne semble pas provenir de l'approximation du processus de Wright-Fisher par un processus de diffusion. En effet, pour $N = 100000$, et en modifiant simultanément t et c de manière à ce que τ et ρ restent les mêmes que dans la figure 6.1, la courbe de vraisemblance obtenue par Monte Carlo reste la même. Dans ce cas, l'erreur vient donc probablement de notre résolution approchée de l'équation de Kolmogorov. Au cours du chapitre précédent, nous avons vu que notre méthode était précise pour des temps d'évolution τ entre 0.01 et 0.1. Le temps $\tau = 0.005$ est peut être trop faible ici. Le fait de prendre une fréquence initiale au bord du domaine ($\pi_{Q,2} = 0$) est peut être aussi un désavantage pour notre méthode, bien que les conditions aux bords soient mieux définies ici que dans le problème du chapitre précédent. Notons pour finir que quand une des composantes du vecteur d'observations n est strictement inférieure à 3, le maximum de la vraisemblance $\mathcal{L}^{(KBE)}$ est toujours en $c = 0$ (ceci a été vérifié pour différentes valeurs de τ, π^0 et n). Ceci pourrait aussi être dû à un problème numérique.

Nous avons également comparé les différentes vraisemblances pour des valeurs des paramètres plus proches de celles typiquement rencontrées en génétique animale, par exemple $N = 250$ et $t = 50$. Pour de telles valeurs de N, la variabilité dans l'évolution des fréquences est beaucoup plus importante, et il est assez difficile d'obtenir une courbe de vraisemblance stable par Monte Carlo. En utilisant 2000 réplicats par valeur de c, nous avons obtenu la courbe approximative représentée par la figure 6.2. Bien que N soit plus faible que dans l'exemple précédent, la courbe de vraisemblance obtenue par diffusion est assez proche de celle obtenue par Monte Carlo, ce qui peut être dû à la différence d'échelle de temps entre les deux exemples : $\tau = 0.005$ pour la figure 6.1, et $\tau = 0.1$ pour la figure 6.2. En revanche, l'approximation d'ordre 1 est beaucoup plus fausse ici que dans l'exemple précédent, et n'admet plus le même maximum. Elle n'est pas représentée sur la figure 6.2.

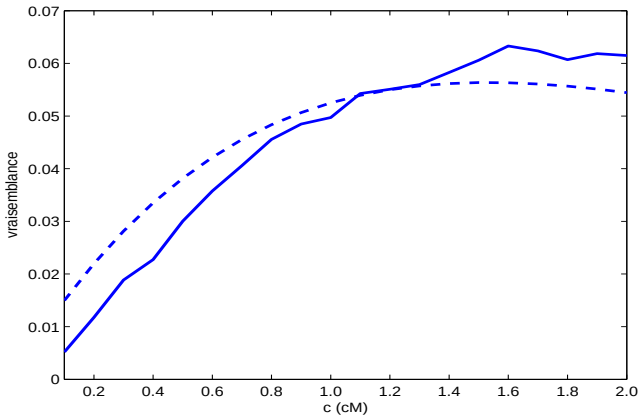
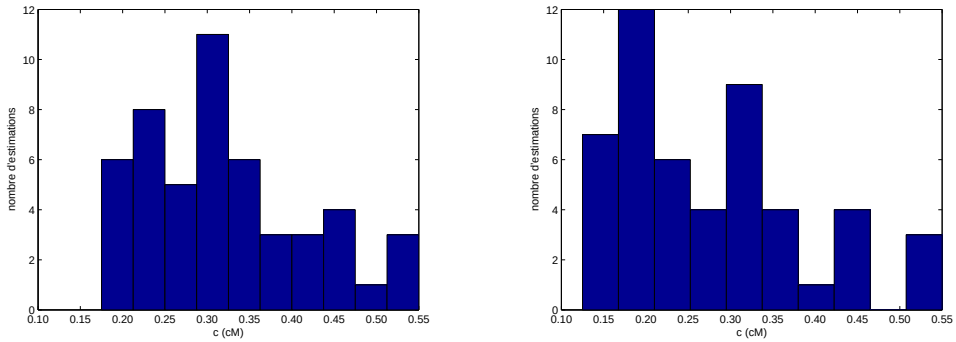


FIG. 6.2 – Courbes de vraisemblance $\mathcal{L}^{(MC)}(c, t, N, \pi^0 \mid n)$ (ligne continue) et $\mathcal{L}^{(KBE)}(\rho, \tau, \pi^0 \mid n)$ (ligne en pointillés) en fonction de c. On a $N = 250$, $t = 50$, $\pi^0 = (0.21, 0, 0.37)$ et $n = (30, 5, 65, 50)$. Pour $\mathcal{L}^{(MC)}(c, t, N, \pi^0 \mid n)$ on simule $M = 2000$ populations pour chaque valeur de c.

6.4 Résultats de simulations

Les estimateurs de maximum de vraisemblance basés sur $\mathcal{L}^{(KBE)}$ et $\mathcal{L}^{(1)}$ ont été testés sur un jeu de 50 échantillons de taille $N_s = 75$. La vraisemblance obtenue par Monte Carlo demande trop de temps de calcul et n'a donc pas été considérée. Chaque échantillon a été produit en simulant une population d'haplotypes pour deux loci distants de $0.2cM$. Les simulations ont été conduites sous un modèle de Wright-Fisher de paramètres $N = 10000$, $t = 100$ et $\pi^0 = (0.21, 0, 0.37)$ comme pour la figure 6.1. A l'issue de chaque simulation, l'échantillon était choisi dans les proportions exactes de la population obtenue (l'introduction d'un échantillonage aléatoire des N_s individus a peu d'influence sur les résultats). Les échantillons ayant une coordonnée strictement inférieure à 3 ont été rejetés. Pour tous les autres échantillons, les vraisemblances $\mathcal{L}^{(KBE)}$ et $\mathcal{L}^{(1)}$ étaient évaluées pour différentes valeurs de c comprises entre $0cM$ et $0.6cM$. Comme cela a été expliqué en 6.2, N et t étaient supposés connus mais π^0 était optimisée pour chaque valeur de c.

La figure 6.3 montre les histogrammes des estimations obtenues par les deux méthodes. Leur profil est assez similaire. Le biais est de $0.13cM$ pour l'approximation d'ordre 1 et de $0.07cM$ pour l'approximation par la diffusion, ce qui confirme les observations faites sur la figure 6.1 à propos du décalage entre les 2 courbes. On peut constater qu'aucune de ces deux méthodes ne permet d'obtenir une très bonne estimation de c. Malgré le choix d'une population large et d'un temps d'évolution court, il semble que la variabilité historique soit encore trop importante.



Utilisation de $\mathcal{L}^{(1)}$ Utilisation de $\mathcal{L}^{(KBE)}$

FIG. 6.3 – Histogrammes des estimations de c obtenues par deux méthodes de maximum de vraisemblance. Les paramètres des populations simulées sont $N = 10000$, $t = 100$, $c = 0.2cM$ et $\pi^0 = (0.21, 0, 0.37)$.

6.5 Conclusions

Pour calculer la vraisemblance $\mathcal{L}(c \mid \mathcal{D}) = \mathbb{E}[\mathcal{L}(c \mid \mathcal{D}, \Pi(t))]$, la solution la plus précise est en théorie la méthode de Monte Carlo basée sur un grand nombre de simulations des trajectoires $\Pi(.)$ (cf formule 6.2). Mais cette solution est difficilement applicable car le temps de calcul nécessaire est très élevé. Pour N de l'ordre de 100 ou 1000, $\Pi(.)$ est très variable et le nombre de simulations doit être très important. Pour N de l'ordre de 10000 ou plus, $\Pi(.)$ est moins variable mais chaque simulation demande beaucoup plus de temps.

Pour aller plus vite, on peut utiliser une approximation d'ordre 1 de la vraisemblance comme l'ont proposé Xiong et Guo (1997). On obtient alors l'expresion (6.3), que l'on peut calculer grâce aux résultats connus sur l'espérance des fréquences d'haplotypes dans un modèle de Wright-Fisher. Nous avons vu en 6.3 que pour $N = 10000$, cette approximation peut être efficace pour l'estimation ponctuelle de x, mais donne probablement des intervalles de confiance libéraux. Pour $N = 100$, nos résultats suggèrent que même la précision de l'estimation ponctuelle de x n'est pas garantie.

L'approximation du processus de Wright-Fisher par un processus de diffusion, combinée à l'utilisation d'un schéma numérique de résolution de l'équation rétrograde de Kolmogorov, permet également de calculer une courbe de vraisemblance approchée. Pour les différents cas étudiés ici (N entre 250 et 100000, τ égal à 0.005 ou 0.1), cette approximation paraît assez fiable. Pour l'estimation ponctuelle, elle donne d'aussi bons résultats que l'approximation d'ordre 1 (cf figure 6.3). Par ailleurs elle donne probablement de meilleurs intervalles de confiance. Cependant elle demande un peu plus de temps de calcul, et quelques problèmes ont été constatés, notamment pour des échantillons comprenant

des valeurs inférieures à 3.

Pour vraiment distinguer la part d'erreur qui relève de l'approximation par un processus de diffusion, et celle provenant de notre résolution numérique de l'équation de Kolmogorov, il serait intéressant de comparer $\mathcal{L}^{(KBE)}$ avec une autre méthode de calcul de la vraisemblance basée sur la diffusion. On peut par exemple envisager de simuler un grand nombre de fois l'ancestral recombination graph (cf section 2.2.2) de l'échantillon observé en arrêtant les simulations au temps τ (Bob Griffiths, communication personnelle). C'est une perspective possible pour la suite de ce travail.

Références

Xiong, M., et Guo, S. (1997). Fine scale genetic mapping based on linkage disequilibrium : theory and applications. *Am J Hum Genet*, *60*, 1513-1531.

Chapitre 7

Perspectives

La théorie de la diffusion joue un rôle primordial pour l'étude de l'évolution des fréquences d'haplotypes dans une population. Le chapitre 5 a présenté une méthode numérique permettant de tirer parti de ce cadre d'études, et a donné les conditions dans lesquelles on pouvait l'utiliser. Le chapitre 6 a fourni quant à lui un exemple d'application de cette méthode pour la cartographie génétique. Pour conclure cette partie, des pistes de travaux futurs sur le même thème sont proposées ici. Elles sont liées à des améliorations de la méthode numérique, et à des applications possibles de cette méthode.

7.1 Amélioration de la méthode

Il existe de nombreuses méthodes numériques pour la résolution des équations aux dérivées partielles. Comme cela a été dit en 5.6, une solution serait d'utiliser des méthodes de résolution stochastiques, ce qui reviendrait à simuler le processus de diffusion représentant les fréquences d'haplotypes. Parmi les approches déterministes, d'autres schémas que celui que nous avons adopté peuvent également être essayés, et apportent en théorie une précision supplémentaire. Cependant il ne fait aucun doute que dans notre cas le facteur le plus limitant pour la précision est le problème des conditions aux bords. Quelques améliorations sont suggérées ici.

7.1.1 Conditions aux bords

Pour traiter de manière moins arbitraire le problème des bords, il n'y a pas d'autre solution que de décrire précisément ce que deviennent les trajectoires de fréquences en arrivant au bord. On doit donc introduire de nombreuses fonctions supplémentaires représentant les densités de probabilités sur les bords. Pour résoudre le problème, il faut

ensuite trouver des relations entre ces différentes fonctions.

Notations

Replaçons-nous dans le cadre du chapitre 5. Le vecteur des fréquences d'haplotypes, dans un modèle de diffusion à deux loci bialléliques, est

$$Y(\tau) = (Y_1(\tau), Y_2(\tau), Y_3(\tau)) = (Y_{1,1}(\tau), Y_{1,2}(\tau), Y_{2,1}(\tau))$$

Connaissant le vecteur de fréquences y^0 au temps 0, nous cherchons à calculer la densité de probabilité $f(y_1, y_2, y_3, \tau)$ de $Y(\tau)$ pour $(y_1, y_2, y_3) \in \mathbb{F}_3$. En plus de f, nous avons besoin de différentes fonctions permttant de caractériser la loi de $Y(\tau)$ sur les bords de $\mathbb{F}_3$, à savoir :

- $f_{1,2}(y_1, y_2, \tau)$, densité de $Y(\tau)$ au point $(y_1, y_2, 0)$ par rapport à la mesure de Lebesgues sur $\mathbb{R}^2 \times \{0\}$. Cette fonction est définie sur la face $y_3 = 0$. On définit de même les densités $f_{1,3}$ et $f_{2,3}$ sur les faces $y_2 = 0$ et $y_3 = 0$.
- $f_1(y_1, \tau)$, densité de $Y(\tau)$ au point $(y_1, 0, 0)$ par rapport à la mesure de Lebesgues sur $\mathbb{R} \times \{0\} \times \{0\}$. Cette fonction est définie sur l'arête $y_2 = y_3 = 0$. On définit de même les densités f_2 et f_3 sur les arêtes $y_1 = y_3 = 0$ et $y_1 = y_2 = 0$.
- $P_{0,0,0}(\tau)$, probabilité de l'événement $Y(\tau) = (0, 0, 0)$. On définit de même les probabilités $P_{1,0,0}(\tau)$, $P_{0,1,0}(\tau)$ et $P_{0,0,1}(\tau)$ aux autres coins.

Remarque 4 *Pour résoudre complètement le problème, il faudrait également introduire les desnsités sur les arêtes $y_1 + y_2 = 1$, $y_1 + y_3 = 1$ et $y_2 + y_3 = 1$. Mais cela ne sera pas nécessaire dans ce chapitre, où nous donnons simplement une idée des méthodes à utiliser.*

Probabilités aux coins

Pour calculer la probabilité au coin $(0, 0, 0)$, on peut supposer la relation :

$$\begin{aligned}
\frac{\partial}{\partial \tau} P_{0,0,0}(\tau) &= \lim_{y_1 \to 0} \{\frac{\partial}{\partial y_1}(\frac{1}{2} a_{1,1}^2(y_1, 0, 0) f_1(y_1, \tau)) - b_1(y_1, 0, 0) f_1(y_1, \tau)\} \\
&+ \lim_{y_2 \to 0} \{\frac{\partial}{\partial y_2}(\frac{1}{2} a_{2,2}^2(0, y_2, 0) f_2(y_2, \tau)) - b_2(0, y_2, 0) f_2(y_2, \tau)\} \\
&+ \lim_{y_3 \to 0} \{\frac{\partial}{\partial y_3}(\frac{1}{2} a_{3,3}^2(0, 0, y_3) f_3(y_3, \tau)) - b_3(0, 0, y_3) f_3(y_3, \tau)\} \qquad (7.1)
\end{aligned}$$

Pour comprendre cette relation, considérons d'abord un processus de diffusion stationnaire $Y(\tau)$ en dimension 1, de coefficients infinitésimaux b et a^2. On suppose que ce processus

est à valeurs dans $[l, r]$, où les bords l et r sont absorbants. Pour $l < y < r$, on note $f(y, \tau)$ la densité de $Y(\tau)$ par rapport à la mesure de Lebesgues, et $P_l(\tau)$ et $P_r(\tau)$ les probabilités pour $Y(\tau)$ d'être en l ou en r à l'instant τ. Pour $l < y < r$, on a

$$\mathbb{P}(l \leq Y(\tau) \leq y) = P_l(\tau) + \int_l^y f(z, \tau)\ dz$$

et par conséquent

$$\frac{\partial}{\partial \tau}\mathbb{P}(l \leq Y(\tau) \leq y) = \frac{\partial}{\partial \tau}P_l(\tau) + \frac{\partial}{\partial \tau}\int_l^y f(z, \tau)\ dz$$

Si f est suffisamment régulier, on a aussi

$$\frac{\partial}{\partial \tau}\int_l^y f(z, \tau)\ dz = \int_l^y \frac{\partial}{\partial \tau} f(z, \tau)\ dz$$

et

$$\lim_{y \to l} \int_l^y \frac{\partial}{\partial \tau} f(z, \tau)\ dz = 0$$

On en déduit donc que

$$\frac{\partial}{\partial \tau}P_l(\tau) = \lim_{y \to l} \frac{\partial}{\partial \tau}\mathbb{P}(l \leq Y(\tau) \leq y) \tag{7.2}$$

Pour un processus de diffusion dont les trajectoires sont continues, il est clair que le terme $\frac{\partial}{\partial \tau}\mathbb{P}(l \leq Y(\tau) \leq y)$ ne dépend que des propriétés du processus en y, lesquelles propriétés sont caractérisées par ses coefficients infinitésimaux. En réalité, on a

$$\frac{\partial}{\partial \tau}\mathbb{P}(l \leq Y(\tau) \leq y) = \frac{\partial}{\partial y}(\frac{1}{2}a^2(y)f(y, \tau)) - b(y)f(y, \tau) \tag{7.3}$$

A défaut d'une preuve exacte, on peut donner l'explication intuitive suivante. $b(y)$ étant un terme de dérive, il correspond au déplacement d'un flux de probabilité $b(y)f(y, \tau)$ en direction de r. Par rapport à l'intervalle $[0, y]$, ce flux est sortant si $b(y)$ est positif, ce qui explique le signe moins rencontré dans (7.3). En revanche $a^2(y)$ est un terme de diffusion donc il correspond au déplacement d'un flux de probabilité $\frac{1}{2}a^2(y)f(y, \tau)$ en direction de l, et d'un autre flux de même intensité en direction de r. Si l'intensité du flux est plus forte en $y + \epsilon$ qu'en $y - \epsilon$, où $\epsilon > 0$ et $\epsilon \to 0$, alors l'intervalle $[l, y]$ reçoit plus de flux qu'il n'en donne. Ceci explique le terme $\frac{\partial}{\partial y}(\frac{1}{2}a^2(y)f(y, \tau))$. On conclut finalement des équations

(7.2) et (7.3) que

$$\frac{\partial}{\partial \tau} P_l(\tau) = \lim_{y \to l} \frac{\partial}{\partial y}(\frac{1}{2} a^2(y) f(y,\tau)) - b(y) f(y,\tau) \tag{7.4}$$

On peut vérifier que cette relation se vérifie exactement pour le modèle de diffusion à un locus présenté en 2.1.1. Dans ce cas, $Y(\tau)$ représente la fréquence d'un allèle donné, et on est bien dans le cadre des hypothèses présentées au paragraphe précédent, avec $l = 0$, $r = 1$, $b(y) = 0$ et $a^2(y) = y(1-y)$. On obtient donc

$$\frac{\partial}{\partial \tau} P_0(\tau) = \frac{f(0,\tau)}{2} \tag{7.5}$$

Or il se trouve que les expressions obtenues par Kimura (1955) pour $P_0(\tau)$ et $f(0,\tau)$ vérifient la relation (7.5).

Revenons maintenant au modèle à deux loci qui est l'objet de ce chapitre. Pour obtenir l'équation (7.1), nous supposons simplement que l'augmentation de la probabilité au coin $(0,0,0)$ ne provient que de la diminution de la probabilité sur les 3 arêtes touchant ce coin : $y_2 = y_3 = 0$, $y_1 = y_3 = 0$ et $y_1 = y_2 = 0$. On applique donc la partie droite de l'équation (7.4) à chaqune de ces 3 arêtes. Cette démarche a déjà été utilisée par Tier et Keller (1978) pour étudier un processus de diffusion en dimension 2, également dans le contexte de processus génétiques. En utilisant les expressions connues de $a^2(y)$ et $b(y)$, l'équation (7.4) donne finalement

$$\frac{\partial}{\partial \tau} P_{0,0,0} = \frac{f_1(0,\tau) + f_2(0,\tau) + f_3(0,\tau)}{2}$$

Densité sur les arêtes

Ce type de raisonnement peut être étendu pour obtenir une relation sur les densités des arêtes. Prenons par exemple l'arête $y_1 = y_3 = 0$. On suppose que la variation de $f_2(y_2,\tau)$ est due à la variation de la densité sur les 2 faces voisines ($y_1 = 0$ et $y_3 = 0$). Mais il faut aussi tenir compte ici de la variation "interne" sur l'arête. Ce qui donne :

$$\begin{aligned}
\frac{\partial}{\partial \tau} f_2 &= \frac{\partial^2}{\partial y_2^2}(a_{2,2}(0,y_2,0) f_2) - \frac{\partial}{\partial y_2}(b_2(0,y_2,0) f_2) \\
&+ \lim_{y_1 \to 0} \{ \frac{\partial}{\partial y_1}(a_{1,1}(y_1,y_2,0) f_{1,2}) + \frac{\partial}{\partial y_2}(a_{1,2}(y_1,y_2,0) f_{1,2}) - b_1(y_1,y_2,0) f_{1,2} \} \\
&+ \lim_{y_3 \to 0} \{ \frac{\partial}{\partial y_2}(a_{2,3}(0,y_2,y_3) f_{2,3}) + \frac{\partial}{\partial y_3}(a_{3,3}(0,y_2,y_3) f_{2,3}) - b_3(0,y_2,y_3) f_{2,3} \}
\end{aligned}$$

où la première ligne représente la variation sur l'arête $y_1 = y_3 = 0$, la deuxième ligne représente la variation sur la face $y_3 = 0$, et la troisième ligne représente la variation sur la face $y_1 = 0$. Pour alléger les notations, les variables des différentes densités sont implicites dans cette formule. Une équation du même type avait aussi été utilisé par Tier et Keller (1978). On trouve finalement

$$2f_2(y_2,\tau) - \frac{1-2y_2}{2}\frac{\partial}{\partial y_2}f_2(y_2,\tau) + \frac{y_2(1-y_2)}{2}\frac{\partial^2}{\partial y_2^2}f_2(y_2,\tau) = \frac{f_{1,2}(0,y_2,\tau) + f_{2,3}(y_2,0,\tau)}{2}$$

Un raisonnement similaire peut permettre d'obtenir une relation entre la limite de f sur les faces et les densités sur les faces $f_{1,2}$, $f_{1,3}$ et $f_{2,3}$.

Bilan

Malgré ces équations supplémentaires et une meilleure formalisation du problème, on voit qu'il manque encore de l'information pour être capable de tout résoudre proprement. Par exemple, même si on est capable de calculer la densité $f_{1,2}$ sur la face $y_3 = 0$, cela ne nous dit pas ce que vaut la limite de f quand y_3 tend vers 0. Or ce sont ces valeurs limites qui nous manquent pour calculer f. Il faudra trouver d'autres relations entre ces quantités pour pouvoir avancer, comme des équations exprimant la conservation de la probabilité peut-être. On peut aussi penser utiliser des résultats connus issus du modèle à un locus, mais on perdra alors l'aspect général de la méthode : en effet, on n'a par exemple aucun résultat exact pour le modèle à un locus avec sélection.

D'autre part, les équations dérivées ci-dessus sont adaptées à des modèles pour lesquels il y a une probabilité non nulle d'être au bord. La question ne se pose pas pour le modèle à un locus car tous les bords sont absorbants. Mais qu'en est-il ici pour les bords non absorbants, tels que les faces par exemple ? La densité sur ces bords-là doit-elle être uniformément nulle ? Si on fait cette hypothèse, à savoir $f_{1,2} = f_{1,3} = f_{2,3} = 0$, on aboutit facilement aux expressions

$$\begin{aligned} f(0,y_2,y_3,\tau) &= 2\rho y_2 y_3 \\ f(y_1,0,y_3,\tau) &= 2\rho y_1(1-y_1-y_3) \\ f(y_1,y_2,0,\tau) &= 2\rho y_1(1-y_1-y_2) \\ f(y_1,y_2,y_3,\tau) &= 2\rho y_2 y_3 \text{ si } y_1+y_2+y_3 = 1 \end{aligned}$$

J'ai essayé d'intégrer ces conditions à la méthode numérique mais il ne semble pas y avoir d'amélioration significative. Il est d'autre part assez étrange que ces fonctions ne

dépendent pas de τ. Ceci semble donc indiquer que même pour les bords non absorbants il y aurait une densité de probabilité non nulle. La question reste donc ouverte pour l'instant.

Rappelons finalement que si on inclut dans le modèle la possibilité de mutation d'un allèle à un autre, il n'y a plus alors aucune frontière absorbante. On peut sans doute dans ce cas régler le problème de manière un peu plus simple.

D'autres conditions approchées

Sans aller jusqu'à déterminer de manière exacte toutes les conditions aux bords, on peut sans doute adopter des conditions approchées tenant plus compte de la dynamique des fréquences d'haplotypes. Considérons par exemple un point de la face $y_2 = 0$, de coordonnées $(y_1, 0, y_3)$. Pour l'instant on calcule sa densité en annulant la dérivée seconde de f dans la direction orthogonale au plan $y_2 = 0$, ce qui donne

$$f(y_1, 0, y_3, \tau) = 2f(y_1, h, y_3, \tau) - f(y_1, 2h, y_3, \tau),\ h = \frac{1}{I}$$

où I est le nombre d'intervalles de discrétisation. Mais quand on est dans la situation où $Y_{1,2}(\tau) = 0$, la création des haplotypes $(1, 2)$ se fait nécessairement en combinant des haplotypes $(1, 1)$ et des haplotypes $(2, 2)$. Cela a pour effet de nous amener dans la configuration $(y_1 - h, h, y_3 + h)$. Une condition au bord plus judicieuse serait donc peut être

$$f(y_1, 0, y_3, \tau) = 2f(y_1 - h, h, y_3 + h, \tau) - f(y_1 - 2h, 2h, y_3 + 2h, \tau),\ h = \frac{1}{I}$$

Je n'ai pas eu le temps de mettre cette idée en pratique.

7.1.2 Changement de variables

Dans le même ordre d'idées, il peut être intéressant de procéder au changement de variables

$$\left\{ \begin{array}{lcccl} P(\tau) & = & Y_1(\tau) + Y_2(\tau) & = & Y_{1,.}(\tau) \\ Q(\tau) & = & Y_1(\tau) + Y_3(\tau) & = & Y_{.,1}(\tau) \\ D(\tau) & = & Y_1(\tau) - P(\tau)Q(\tau) & & \end{array} \right.$$

et de résoudre l'équation de Kolmogorov pour la nouvelle diffusion ainsi obtenue, dont les paramètres sont

$$a^2(p,q,d) = \begin{pmatrix} p(1-p) & d & d(1-2p) \\ d & q(1-q) & d(1-2q) \\ d(1-2p) & d(1-2q) & pq(1-p)(1-q) + d(1-2p)(1-2q) - d^2 \end{pmatrix}$$

et

$$b(p,q,d) = \begin{pmatrix} 0 \\ 0 \\ \frac{-d(\rho+1)}{2} \end{pmatrix}$$

Cette formulation a l'avantage de présenter clairement le problème sous forme des fréquences alléliques $P(\tau)$ et $Q(\tau)$ et du déséquilibre de liaison $D(\tau)$. L'interprétation des résultats est donc plus intuitive. D'autre part, les conditions aux bords sont peut être plus simples à exprimer car on n'a que 4 arêtes, toutes absorbantes, et deux faces non absorbantes. Une certaine symétrie dans le problème est également restaurée. En revanche, les équations qui définissent les faces sont plus complexes : pour p et q fixés, d varie entre $\min(pq, (1-p)(1-q))$ et $\min(p(1-q), (1-p)q)$.

7.2 Applications

La méthode que nous avons développée est originale car elle permet de caractériser de manière précise la loi des fréquences d'haplotypes en régime transitoire pour un modèle à deux loci liés. De plus, elle peut facilement incorporer des phénomènes de mutation ou de sélection. Jusqu'ici, les seules méthodes permettant de faire la même chose sont basées sur des simulations, ce qui, comme nous l'avons vu, peut être extrêmement long pour des populations de taille importante (de l'ordre de $N \geq 1000$ par exemple). Les applications sont donc nombreuses. En voici quelques exemples.

7.2.1 Mesures de déséquilibre de liaison

Comme le montre l'article de Cierco-Ayrolles et al. (2004) présenté en annexe, l'étude des propriétés des mesures de déséquilibre de liaison est un sujet qui a suscité et suscite encore beaucoup d'intérêt dans la communauté génétique. Parmi les questions les plus classiques, on trouve notamment les suivantes :

- Quelle mesure a la variance la plus faible pour des distances génétiques "courtes" ? Et pour des distances génétiques longues ?

- Jusqu'à quelle distance est-il "normal" d'observer du déséquilibre de liaison pour un modèle sans sélection ?
- Y a-t-il des mesures qui soient plus robustes que d'autres à la présence de sélection ?
- Un taux de mutation fort peut-il avoir le même effet qu'un taux de recombinaison faible sur la structure du déséquilibre de liaison ?

Être capable de déterminer rapidement la loi des fréquences d'haplotypes grâce à la version forward de notre algorithme devrait permettre d'étudier de très nombreux cas de figure et d'apporter des réponses à ces questions.

7.2.2 Estimation de paramètres

Nous avons vu au chapitre 6 un exemple d'application de la version backward de notre algorithme pour l'estimation du taux de recombinaison entre deux loci à partir de l'observation d'un échantillon d'individus issus de la génération actuelle. De manière générale, il existe de nombreuses méthodes pour estimer les paramètres génétiques que sont le taux de recombinaison, de sélection, de mutation, de migration, ...La plupart d'entre elles utilisent des algorithmes MCMC ou d'importance sampling basés sur les arbres de coalescence des individus observés (voir section 4.3.3). Notre méthode n'est pas forcément plus rapide que ces dernières car il faut lancer une résolution numérique pour chaque valeur des paramètres génétiques (encore que cette opération soit très facilement parallélisable), mais elle présente tout de même deux particularités intéressantes :

- Elle permet l'estimation conjointe du taux de recombinaison et du taux de sélection, ce qui assez délicat en utilisant des arbres de coalescence. D'ailleurs aucune méthode ne le propose actuellement.
- Elle est très adaptée aux cas où on dispose de plusieurs échantillons issus de générations différentes, puisqu'elle calcule des densités de transition. Or des données de ce type sont parfois publiées en génétique, par exemple quand des animaux fossiles sont découverts (Mark Beaumont, communication personnelle).

La version forward de notre algorithme n'est quant à elle pas directement utilisable pour l'analyse de données. En revanche, de nombreuses méthodes d'estimation Bayesiennes ont recours à des distributions a priori pour les fréquences d'haplotypes dans la population. Ces distributions sont souvent non informatives ou, parfois, correspondent à des distributions stationnaires pour des loci indépendants. Si on a des connaissances par rapport au temps d'évolution, à la distance entre loci, ..., l'utilisation de notre algorithme pourrait fournir une bonne distribution a priori.

Références

Cierco-Ayrolles, C., Abdallah, J., Boitard, S., Chikhi, L., Rochambeau, H. de, Tsitrone, A., et al. (2004). On linkage disequilibrium measures : Methods and applications. In (Vol. 1, p. 151-180). Research Signpost, India.

Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Nat. Acad. Sci. USA*, *41*, 144-150.

Tier, C., et Keller, J. B. (1978). Asymptotic analysis of diffusion equations in population genetics. *SIAM J. Appl. Math.*, *34* (3), 549-576.

Troisième partie

Cartographie de QTL par interval mapping

Chapitre 8

Article : Linkage disequilibrium interval mapping of quantitative trait loci

Résumé en français

Au cours du chapitre 6, nous avons pu comparer différentes méthodes pour calculer la vraisemblance de la position x d'un gène. Les données utilisées pour le calcul en chaque position x étaient issues d'un seul marqueur. Pour améliorer la précision des estimations, il semble nécessaire d'analyser conjointement l'information de plusieurs marqueurs. L'article présenté dans ce chapitre propose une méthode d'estimation de x dans laquelle le calcul de la vraisemblance en un point x utilise l'information des deux marqueurs *flanquants*, c'est à dire des deux marqueurs les plus proches de la position, l'un étant à gauche et l'autre à droite de cette position (voir figure 8.1).

Comme au chapitre 6, la vraisemblance en x fait intervenir le vecteur caché $\Pi(t)$ des fréquences d'haplotypes dans la population, soit

$$\mathcal{L}(x \mid \mathcal{D}) = \mathbb{E}[\mathcal{L}(x \mid \mathcal{D}, \Pi(t)]$$

Mais cette fois les haplotypes dont les fréquences sont représentées par $\Pi(t)$ correspondent à un système de trois loci comprenant les deux marqueurs et le QTL de position présumée x. Nous utilisons une approximation d'ordre 1 de la vraisemblance, ce qui nécessite de connaître l'expression de $\mathbb{E}[\Pi(t)]$ pour un modèle de Wright-Fisher à trois loci liés. Xiong et Guo (1997) ont fourni une telle expression dans le cas ou le gène à localiser a deux allèles dont un de fréquence très rare. Nous étendons leurs calculs et obtenons une expression générale pour un gène d'intérêt biallélique. Nous supposons toutefois -en justifiant pourquoi- que les fréquences marginales des allèles aux marqueurs sont constantes.

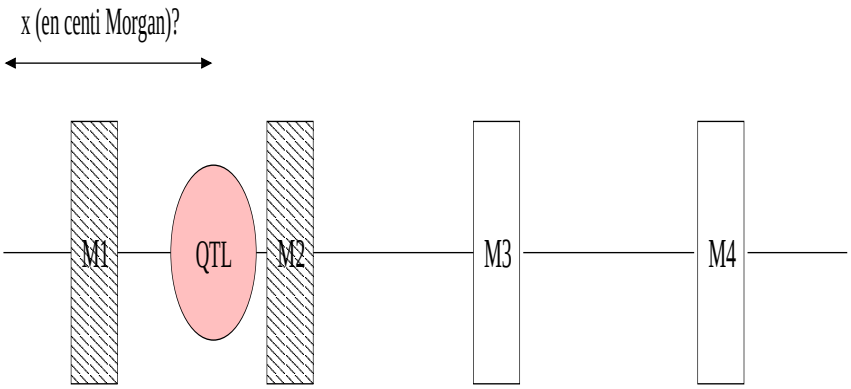


FIG. 8.1 – Les deux marqueurs hachurés (M1 et M2) sont les marqueurs flanquants pour le QTL quand il est dans cette position de la carte.

Le modèle de relation génotype / phénotype est un peu plus évolué que celui considéré au chapitre 6 : c'est un modèle de mélange, dans lequel le phénotype d'un individu est issu d'une loi normale dont la moyenne dépend du génotype au QTL. Ce type de modèles a été décrit en 1.4 dans le cas général de plusieurs QTL. Ici on suppose qu'il y a un seul QTL dans la zone considérée. D'autre part, le calcul de la vraisemblance en chaque position x est très rapide. Contrairement aux méthodes étudiées au chapitre 6, la méthode présentée dans ce chapitre est donc directement utilisable en pratique pour un échantillon d'individus non apparentés pour lesquels les haplotypes aux marqueurs sont connus.

Sur plusieurs scénarios de simulations, nous avons calculé les erreurs quadratiques moyennes sur l'estimation de x obtenue par notre méthode ainsi que par d'autres méthodes existantes. Ces résultats montrent que notre méthode est plus précise que la méthode d'Abdallah et al (2004), qui calcule la vraisemblance en x en multipliant toutes les vraisemblances obtenues à partir d'un seul marqueur. Notre méthode est également aussi précise que la méthode IBD de Meuwissen et Goddard (2000) qui a été décrite en 4.3.2 quand celle-ci est utilisée avec deux marqueurs.

Remarque 5 *Nous présentons ici l'article sous la forme définitive adoptée par la revue BMC Genomics. Il y a donc quelques différences de notations par rapport aux chapitres précédents. Celles-ci sont présentées dans le tableau 8.1. Les notations de ce chapitre sont par ailleurs étendues au reste de la partie III.*

Variable	Notation partie III	Notation autres parties
Phénotypes	Y	Z
Probabilités de transition*	$r(t)$	p
Fréquences alléliques (marqueur 1)	Π_{i_1}	$\Pi_{i_1,.}$
Fréquences alléliques (marqueur 2)	Π_{i_2}	$\Pi_{.,i_2}$

* : pour le modèle de Wright-Fisher

TAB. 8.1 – Différences de notations entre la partie III et le reste du manuscrit

BMC Genomics

BioMed Central

Methodology article

Open Access

Linkage disequilibrium interval mapping of quantitative trait loci

Simon Boitard*[1,2], Jihad Abdallah[3,4], Hubert de Rochambeau[4], Christine Cierco-Ayrolles[1,2] and Brigitte Mangin[1]

Address: [1]Unité de Biométrie et Intelligence Artificielle, Institut National de la Recherche Agronomique, BP 52627, 31326 Castanet-Tolosan Cedex, France, [2]Laboratoire de Statistiques et Probabilités, Université Paul Sabatier, 118 route de Narbonne, 31400 Toulouse, France, [3]Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, BP 52627, 31326 Castanet-Tolosan Cedex, France and [4]Station d'Amélioration Génétique des Animaux, Institut National de la Recherche Agronomique, BP 52627, 31326 Castanet-Tolosan Cedex, France

Email: Simon Boitard* - simon.boitard@toulouse.inra.fr; Jihad Abdallah - jihad.abdallah@toulouse.inra.fr; Hubert de Rochambeau - rochambeau@toulouse.inra.fr; Christine Cierco-Ayrolles - christine.cierco@toulouse.inra.fr; Brigitte Mangin - brigitte.mangin@toulouse.inra.fr

* Corresponding author

Published: 16 March 2006

BMC Genomics 2006, **7**:54 doi:10.1186/1471-2164-7-54

Received: 04 November 2005
Accepted: 16 March 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/54

Abstract

Background: For many years gene mapping studies have been performed through linkage analyses based on pedigree data. Recently, linkage disequilibrium methods based on unrelated individuals have been advocated as powerful tools to refine estimates of gene location. Many strategies have been proposed to deal with simply inherited disease traits. However, locating quantitative trait loci is statistically more challenging and considerable research is needed to provide robust and computationally efficient methods.

Results: Under a three-locus Wright-Fisher model, we derived approximate expressions for the expected haplotype frequencies in a population. We considered haplotypes comprising one trait locus and two flanking markers. Using these theoretical expressions, we built a likelihood-maximization method, called HAPim, for estimating the location of a quantitative trait locus. For each postulated position, the method only requires information from the two flanking markers. Over a wide range of simulation scenarios it was found to be more accurate than a two-marker composite likelihood method. It also performed as well as identity by descent methods, whilst being valuable in a wider range of populations.

Conclusion: Our method makes efficient use of marker information, and can be valuable for fine mapping purposes. Its performance is increased if multiallelic markers are available. Several improvements can be developed to account for more complex evolution scenarios or provide robust confidence intervals for the location estimates.

Background

The detection and mapping of loci affecting quantitative traits (QTLs) of interest in human, animal, and plant populations have attracted considerable research interest for several decades. This work has mainly concentrated on the use of pedigree or family data, especially in animal and plant populations where the structure of such experimental pedigrees can easily be planned and controlled. However, it is difficult to attain an accuracy of less than 5 centimorgans (cM) for the gene locations estimated by such linkage analysis methods because of the small

number of meioses occurring in only a few generations [1,2].

More recently, linkage disequilibrium (LD) methods based on the study of unrelated individuals from a given population have emerged as a promising tool for refining gene location estimates. These methods are based on the following key hypothesis [3,4]: when a new allele is introduced into a population, either by mutation or migration, it exists in that population with a unique set of marker alleles. The length of this characteristic haplotype is then reduced along generations by recombination events, and after many generations only the markers in the immediate vicinity of the new allele locus are likely to remain on the same strand. If the new allele has a particular influence on a given trait, a strong correlation between this trait value and a marker allele might thus indicate that the coding locus is very close to the marker.

In practice, the earlier successes in mapping genes using such strategies concerned simply inherited (Mendelian) disease genes in isolated human populations [3,5-7], and the many mapping methods that have been subsequently developed for this kind of problem can be roughly divided into two classes: (i) forward analyses of allele or haplotype frequencies in the disease (case) and normal (control) populations [8-13], and (ii) backward inferences of the case sample genealogy using coalescence [14-16]. Some of these methods are specifically designed for populations divided into cases and controls, and take advantage of the assumption that the allele responsible for the disease is rare. Consequently, they are difficult to extend to mapping QTLs or complex disease traits.

The association between a quantitative trait and a marker allele can be exploited in QTL mapping. This was first proposed in [17] through a simple analysis-of-variance framework. We [18] and Farnir and colleagues [19] subsequently used a maximum-likelihood approach, based on the same kind of allele frequency model as in [9] but for the purpose of QTL mapping. Pérez-Enzizo [20] provided a method based on a hidden Markov model for marker identity by descent (IBD) with the ancestral haplotype [13]. Meuwissen and Goddard [21,22] integrated the LD information in a mixed linear model through a matrix of IBD probabilities for the sample marker haplotypes. They used the so-called gene-dropping method and approximate theoretical expressions to compute these probabilities. More recently, Zöllner and Pritchard [23] developed a Bayesian method based on backward simulations of the sample ancestry using a local approximation of the ancestral recombination graph [24]. Encouraging results were also obtained in practice. For instance, an allele substitution that has a major effect on milk yield and composition was identified using LD information [25]. The present interest in finding new associations is fuelled by the increasing number of new polymorphic markers available on human and livestock genomes. However, QTL mapping remains a statistical challenge due to the weak phenotype-genotype correlation and the influence of environmental or multigene factors. Furthermore, the accuracy and computational efficiency of mapping methods still need to be increased.

Our method is an interval-mapping method designed for unrelated individuals with no family information, and is based on a maximum-likelihood calculation. Computations of the likelihood function at each postulated location of the QTL rely on the expected frequencies of a three-locus haplotype composed of the QTL and its two flanking markers. We provide an approximate expression of these expected frequencies at time *t*, assuming a Wright-Fisher model for the population and a punctual creation of LD at time 0, as described above. Due to this approximation the computation time required by our method is very low.

In this paper, we first describe the model we use and explain the differences between our method and existing ones. We then report the results of a simulation study, in which we test our method under various evolution scenarios, and compare it with the composite two-marker method in [18] and the multimarker methods in [21,26,27]. Finally we discuss the advantages and drawbacks of our method, as well as the potential improvements that could be implemented.

Results

Maximum likelihood approach

We consider a single quantitative trait whose value is partly controlled by a biallelic locus with alleles *Q* and *q*. As usual (and following [28]), the probability density of phenotype *Y* conditional on QTL genotype **G** is modeled as follows:

$$\mathrm{d}\mathbb{P}(Y = y \mid \mathbf{G}) = \begin{cases} \phi_{\mu+a,\ \sigma^2}(y) & \text{if } \mathbf{G} = Q/Q \\ \phi_{\mu+a,\ \sigma^2}(y) & \text{if } \mathbf{G} = q/q \\ \phi_{\mu+d,\ \sigma^2}(y) & \text{if } \mathbf{G} = Q/q \end{cases} \qquad (1)$$

where ϕ_{m,σ^2} is the density function of a normal distribution $\mathcal{N}(m, \sigma^2)$ *a* is the additive effect of the QTL, *d* is the dominance effect, and μ is the mean trait value for homozygotes.

Our data contain N_s unrelated individuals sampled from the same population. We observe their phenotypic values

BMC Genomics 2006, **7**:54 http://www.biomedcentral.com/1471-2164/7/54

y_n, $n = 1,\ldots, N_s$, and their genotypes $\mathbf{m}_n$ for a given set of markers. For the purpose of generality, we do not yet specify how many of these markers there are. Our aim is to estimate as accurately as possible the position x of the QTL on the known marker map, for which we use a multipoint approach consisting of computing – for a large number of positions x of the QTL – the likelihood function $\mathcal{L}(x \mid \mathcal{D})$, where $\mathcal{D} = \{(y_n, \mathbf{m}_n), n = 1,\ldots, N_s\}$. The value of x that maximizes this likelihood function will be the estimate of the QTL position.

Since individuals are unrelated, the pairs of random variables $(Y_n, \mathbf{M}_n)$ can be considered as independent. Therefore, the likelihood function is

$$\mathcal{L}(x \mid \mathcal{D}) = \prod_{n=1}^{N_s} \mathbb{P}(Y_n = y_n, \mathbf{M}_n = \mathbf{m}_n \mid x) \propto \prod_{n=1}^{N_s} \mathbb{P}(Y_n = y_n \mid \mathbf{M}_n = \mathbf{m}_n, x)$$

where $\propto$ means "proportional to", since the multiplicative constant is independent of x. We exploit the parametric model (1) by deriving the probabilities $\mathbb{P}(Y_n = y_n, \mathbf{M}_n = \mathbf{m}_n \mid x)$, $n = 1,\ldots, N_s$, conditional on the random variables $\mathbf{G}_n$ that denote the QTL genotype for individual n. We get for all n that

$$\begin{aligned}\mathbb{P}(Y_n = y_n \mid \mathbf{M}_n = \mathbf{m}_n, x) &= \phi_{\mu_{Q/Q}, \sigma^2}(y_n)\ \mathbb{P}(\mathbf{G}_n = Q/Q \mid \mathbf{M}_n = \mathbf{m}_{n,x}) \\ &+ \phi_{\mu_{Q/q}, \sigma^2}(y_n)\mathbb{P}(\mathbf{G}_n = q/q \mid \mathbf{M}_n = \mathbf{m}_{n,x}) \\ &+ \phi_{\mu_{q/q}, \sigma^2}(y_n)\mathbb{P}(\mathbf{G}_n = Q/q \mid \mathbf{M}_n = \mathbf{m}_{n,x})\end{aligned}$$

Let us now assume that the haplotype phases are known. Each genotype $\mathbf{m}_n$ can thus be written as the diplotype $\mathbf{h}_n^1 / \mathbf{h}_n^2$, where $\mathbf{h}_n^1$ and $\mathbf{h}_n^2$ belong to the set of all haplotypes that can be found in the population for the L marker loci. Let j_n^1 and j_n^2 be their respective indexes in this set. For any haplotype $\mathbf{h}$ of index j, we denote Π_j as its frequency in the population and $\Pi_{Q,j}$ as the frequency of haplotype $(Q, \mathbf{h})$ in the population. Conditionally on the vector Π of all haplotype frequencies in the population and assuming Hardy-Weinberg equilibrium, we can now express the probabilities of QTL genotypes given the marker genotypes as follows:

$$\begin{aligned}\mathcal{L}(x \mid \Pi) \propto \prod_{n=1}^{N_s}\Bigg[&\phi_{\mu_{Q/Q}, \sigma^2}(y_n)\frac{\Pi_{Q,j_n^1}}{\Pi_{j_n^1}}\frac{\Pi_{Q,j_n^2}}{\Pi_{j_n^2}} + \phi_{\mu_{Q/q}, \sigma^2}(y_n)\frac{\Pi_{q,j_n^1}}{\Pi_{j_n^1}}\frac{\Pi_{q,j_n^2}}{\Pi_{j_n^2}} \\ &+ \phi_{\mu_{q/q}, \sigma^2}(y_n)\left(\frac{\Pi_{q,j_n^1}}{\Pi_{j_n^1}}\frac{\Pi_{Q,j_n^2}}{\Pi_{j_n^2}} + \frac{\Pi_{Q,j_n^1}}{\Pi_{j_n^1}}\frac{\Pi_{q,j_n^2}}{\Pi_{j_n^2}}\right)\Bigg]\end{aligned} \quad (2)$$

However, the haplotype frequencies in the population are random variables evolving stochastically along generations, and their values at the time that the data are sampled are unknown. Thus the true likelihood is

$$\mathcal{L}(x \mid \mathcal{D}) = \mathbb{E}[\mathcal{L}(x \mid \mathcal{D}, \Pi)] \quad (3)$$

where the expected value is taken over the probability distribution of haplotype frequencies in the population. This distribution depends on parameters such as the effective population size and the recombination rates between loci, and is specified by mathematical models of population genetics. The general idea of computing the likelihood conditionally on haplotype frequencies in the population and then taking the expected value was first proposed in [29], and was subsequently used in [10] and [8]. However, all these papers were dealing with dichotomous disease traits for which the form of the likelihood was quite different.

Approximating the likelihood

Under classical models of population genetics, the likelihood function defined by (2) and (3) cannot be easily calculated, and so approximations are necessary. A natural approach is to estimate (3) using a Monte Carlo method, simulating a large number of population replicates for one marker and one disease gene [8]. Unfortunately this approach is very time consuming. In fact, a huge proportion of replicates have to be dropped because the allele frequencies at the final generation are not in good agreement with the ones observed in the sample. A more direct way of computing (3) is to approximate the overall expected value by a expected values; i.e.,

$$\begin{aligned}\mathcal{L}(x \mid \mathcal{D}) = \prod_{n=1}^{N_s}\Bigg[&\phi_{\mu_{Q/Q}, \sigma^2}(y_n)\frac{\mathbb{E}[\Pi_{Q,j_n^1}]}{\mathbb{E}[\Pi_{j_n^1}]}\frac{\mathbb{E}[\Pi_{Q,j_n^2}]}{\mathbb{E}[\Pi_{j_n^2}]} + \phi_{\mu_{Q/q}, \sigma^2}(y_n)\frac{\mathbb{E}[\Pi_{Q,j_n^1}]}{\mathbb{E}[\Pi_{j_n^1}]}\frac{\mathbb{E}[\Pi_{q,j_n^2}]}{\mathbb{E}[\Pi_{j_n^2}]} \\ &+ \phi_{\mu_{q/q}, \sigma^2}(y_n)\left(\frac{\mathbb{E}[\Pi_{q,j_n^1}]}{\mathbb{E}[\Pi_{j_n^1}]}\frac{\mathbb{E}[\Pi_{Q,j_n^2}]}{\mathbb{E}[\Pi_{j_n^2}]} + \frac{\mathbb{E}[\Pi_{Q,j_n^1}]}{\mathbb{E}[\Pi_{j_n^1}]}\frac{\mathbb{E}[\Pi_{q,j_n^2}]}{\mathbb{E}[\Pi_{j_n^2}]}\right)\Bigg]\end{aligned} \quad (4)$$

As a consequence of Taylor's expansion and convergence in probability of Π, (4) can be proved to converge to the true likelihood as the effective population size tends to infinity. Using this formula is equivalent to assuming that the effective population size is infinite, or that changes in haplotype frequencies along generations are deterministic. This approximation can be refined by adding the second term of the Taylor expansion, which involves second moments of haplotype frequencies $\Pi_{Q,j}$. This was done in the context of a single-marker method by Xiong and Guo [10], who concluded that introducing this second-order term did not significantly improve the location estimates. Therefore, in the following sections we focus on methods using only the first-order approximation in (4).

Mixture model

Using approximation (4), our model can be described as follows. Each phenotype value Y_n is randomly drawn from the mixture of three normal distributions: $\phi_{\mu=a,\ ^2}$, $\phi_{\mu=a,\ ^2}$ and $\phi_{\mu=d,\ ^2}$. The probabilities of being drawn from each of these distributions result from the genetic history of the population. They can be derived under a few assumptions on the population model, as illustrated in the following sections. These probabilities depend on the diplotype $\mathbf{h}_n^1 / \mathbf{h}_n^2$. At the first order, our method is thus equivalent to fitting a linear model $\mathbf{Y} = \mathbf{X}\theta + \varepsilon$, where $\mathbf{Y}$ is the vector of phenotype records, θ is the vector of diplotype effects, ε is a vector of independent random noises with variance σ^2 and $\mathbf{X}$ is a design matrix of size $N_s \times D$, D being the number of diplotypes in the population. Each component of θ is a known function of a small number of population parameters which model the LD creation and the evolution process of the population. Each component of θ is also supposed to fit the phenotype mean observed for one particular diplotype, so that each diplotype provides one equation. Our aim is to identify the population parameter values that are optimal with respect to the whole set of equations.

Using marker information

The simultaneous use of more markers should increase the accuracy of the QTL location because the past recombination events can be identified more precisely. However, increasing the number of markers makes the computation of haplotype frequency distribution – and consequently of the likelihood function in (4) -more complex. We previously [18] provided two methods for fine mapping of quantitative traits. The first one was a single-marker method: for each position x on the map, only one marker was considered and the expected haplotype frequencies $\mathbb{E}[\prod_{Q,i}]$ and $\mathbb{E}[\prod_{q,i}]$ were expressed for every allele i of this marker as a function of the allele frequencies, the time t since the initial creation of LD, the recombination rate c between the QTL and the marker locus, the allele initially associated with the mutation Q, and a heterogeneity parameter α that is described in more detail below. Equation (4) could thus be computed. With only one marker, parameters t, c and α could not be estimated independently of each others so they were combined into a single parameter $\lambda = \alpha(1 - c)^t$. The second method was a composite likelihood method that used the set of L closest markers at each position whilst assuming that these markers were associated with the QTL independently of each other:

$$\mathcal{L}(x \mid \mathcal{D}) = \prod_{l=1}^{L} \mathcal{L}_l(x \mid \mathcal{D})$$

where $\mathcal{L}_l(x \mid \mathcal{D})$ denotes the single-marker likelihood function for the lth marker.

The above assumption of independence is clearly violated when markers are linked. To account for a correlation between close loci, Xiong and Guo [10] determined an expression for the expected frequency of haplotypes with one disease gene and two markers. They computed the likelihood function (4) using – at each postulated position of the disease locus – the information from the two flanking markers. Their method takes into account recurrent mutations and population growth since the initial creation of LD. For several experimental data sets, Xiong and Guo showed that their method provided better estimations than those in [8] and [9]. However, their method is based on the assumption that the allele causing the disease is rare, which allows the haplotype frequencies in the healthy population to be modeled as a deterministic process and thus simplifies the derivations.

The above assumption is not appropriate when dealing with QTLs. Consequently, we extended the derivations in [10] to the general case where all haplotype frequencies are random variables following the three-locus Wright-Fisher model. The allele frequencies at markers are still assumed to be deterministic, time invariant, and in equilibrium in the sense that if i_1 and i_2 respectively denote alleles of the left- and right-side markers, $\Pi_{i_1,i_2} = \Pi_{i_1}\Pi_{i_2}$. We proved that the expected frequency of haplotype (i_1, Q, i_2) after t generations is given by

$$\begin{aligned}\mathbb{E}[\Pi_{i_1,Q,i_2}(t)] = {} & \Pi_Q(0)\Pi_{i_1}\Pi_{i_2} + (1-c_1)^t(\Pi_{i_1,Q}(0) - \Pi_Q(0)\Pi_{i_1})\Pi_{i_2} + (1-c_2)^t(\Pi_{Q,i_2}(0) - \Pi_Q(0)\Pi_{i_2})\Pi_{i_1} \\ & + (1-c_1)^t(1-c_2)^t(\Pi_{i_1,Q,i_2}(0)\Pi_{i_1,Q}(0) - \Pi_{i_2} - \Pi_{Q,i_2}(0)\Pi_{i_1} + \Pi_Q(0)\Pi_{i_1}\Pi_{i_2})\end{aligned} \quad (5)$$

where c_1 and c_2 respectively denote the recombination rates with the left- and right-side markers, and $\Pi_{i_1,Q,i_2}(0), \Pi_{i_1,Q}(0), \Pi_{Q,i_2}(0)$, and $\prod_Q(0)$ are the frequencies of haplotypes (i_1, Q, i_2), (i_1, Q), and (Q, i_2), and allele Q at generation 0, respectively. The derivation of this formula is given in the Appendix. At each postulated location x, c_1 and c_2 are deduced from the marker map and the expected value (5) can be used to compute the likelihood (4)

Initial creation of LD

Our method relies on the assumption that the haplotype frequencies in the population were in equilibrium until a genetic or demographic event suddenly created LD between the QTL and a unique marker haplotype at time 0. Classical examples of such events are the introduction of a favorable allele *Q* into an isolated population, by mutation or migration. After this event, haplotype frequencies evolve along generations as described by (5) until the present generation denoted as *t*.

This model allows us to reduce the number of parameters used to describe haplotype frequencies at time 0. Indeed, following [9] and [10], we introduce a heterogeneity parameter α in addition to allele frequencies Π_{i_1}, Π_{i_2}, and $\Pi_Q(0)$. This parameter represents the proportion of new copies of allele *Q* introduced at time 0 into the population. Note that $\alpha = 1$ if *Q* did not exist previously in the population. Assuming that new alleles *Q* are associated with allele 1 of both markers, the initial frequencies of (5) can be expressed as

$$\Pi_{i_1,Q}(0) = (1-\alpha)\Pi_{i_1}\Pi_Q(0) + \alpha\Pi_Q(0)\delta_{i_1=1}$$
$$\Pi_{Q,i_2}(0) = (1-\alpha)\Pi_{i_2}\Pi_Q(0) + \alpha\Pi_Q(0)\delta_{i_2=1}$$
$$\Pi_{i_1,Q,i_2}(0) = (1-\alpha)\Pi_{i_1}\Pi_{i_2}\Pi_Q(0) + \alpha\Pi_Q(0)\delta_{i_1=1}\delta_{i_2=1}$$

where $\delta_{x=y}$ is the Kronecker delta operator (equal to 1 if $x = y$ and 0 otherwise).

This model can even be used in a more general context than the introduction of a new allele into an isolated population. Indeed, we know that many of the current isolated populations in both humans and animals [30,31] were initially created by a severe bottleneck in a wider population, implying the underrepresentation of many haplotypes and the over represent at ion of others. After such events, it would not be surprising for an allele of rather low frequency to become associated in the new population with a very small number of marker haplotypes. Our model thus applies to that case, provided that time 0 refers to the creation of the population (while the mutation occurred earlier). Parameter α then represents the excess of the overrepresented haplotype including allele *Q*. However, this is only a rough approximation since the favorable allele may in general be associated with more than one haplotype. Many animal breeding populations have also been created by the artificial admixing of two other populations (see [31] for a review), but the amount of LD created between two loci depends on the difference of allele frequencies at these loci between the initial populations. Since this difference is not the same for all loci, there is no reason why a single unique coefficient α should be used to model the initial level of association of *Q* with all markers. Consequently, our method appears to be unsuitable for such cases.

Simulation Results

As outlined above, one fundamental feature of a mapping method is its ability to simultaneously use the information from several markers. We have previously [18] proposed a single-marker method (Tl) and two composite likelihood methods (T2 and T6) to map QTLs using LD. Based on simulation results, our conclusions were that (i) composite likelihood methods provide better location estimates than single-marker methods such as regression analysis or Tl, and (ii) among composite likelihood methods, the one using two markers (T2) generally performs the best.

Starting from these conclusions, we first compare our new method – which we have called HAPim – with T2. While haplotype methods are generally considered to be more accurate than composite likelihood ones, we considered it important to evaluate the exact difference between them, as well as the influence of parameters such as effective population size, marker spacing, and time since the initial creation of LD. We also discuss the behavior of both methods in the presence of incomplete association or phenocopies. We then compare the accuracy of our method with that of the haplotype method in [21]. Both of the following analyses are based on the simulation framework described in the Methods section.

Comparison with a composite likelihood method

We first compared HAPim and T2 by reproducing simulation scenarios similar to those in [18]. The QTL was simulated at position 3.6 cM on a 10-cM marker map. Two effective population sizes (N = 200 and N = 400), two marker-spacing values (0.25 and 2 cM), and both single nucleotide polymorphisms (SNPs) and microsatelites (MSTs) were tested. The time since the initial LD creation was t = 100, and no copy of allele *Q* was present in the population before that time, which ensured that complete initial LD was present. The mean square errors (MSEs) of both mapping methods under these various scenarios are given in Table 1. Unsurprisingly, they both performed better with decreasing marker spacing, increasing effective population size and multiallelic markers. However, we were more interested in the influence of parameters on the difference in precision between the methods than on their absolute precisions (which has already been widely studied). Table 1 indicates that the gain from using HAPim is particularly significant with dense maps, irrespective of the marker type and effective population size. This was expected because T2 assumes independence between the QTL-marker associations, which is increasingly violated as the marker spacing decreases.

Table 1: General Comparison between T2 and HAPim.

Marker type	*N*	Marker spacing	MSE		Difference in MSE P value
			T2	HAPim	
SNP(1)	200	2 cM	5.10	5.08	0.948
	400	2 cM	4.69	5.04	0.366
	200	0.25 cM	2.00	1.24	< 0.001**
	400	0.25 cM	1.34	0.92	0.005**
MST(2)	200	2 cM	2.93	2.77	0.438
	400	2 cM	1.81	1.44	0.056
	200	0.25 cM	0.71	0.46	0.012*
	400	0.25 cM	0.49	0.30	0.033*

* : P < 0.05, ** : P < 0.01
(1) : single nucleotide polymorphism
(2) : microsatellite
Mean square errors (MSEs) in cM^2 of quantitative trait locus (QTL) location estimates obtained by the T2 and HAPim methods for various effective population sizes, marker spacings and marker types, t = 100 and the initial association was complete.

Table 2 presents the quality of the estimates for all the model parameters using SNP markers, an effective population size of 400, and a marker spacing of 0.25 cM. The QTL location estimate from HAPim was almost unbiased and, as evident in Table 1, more precise than the one from T2. The additive and dominance effects were also very accurately estimated, again better than T2 for the dominance effect. Both methods slightly underestimated heterogeneity parameter α, due to it being constrained to be less than 1. The time since the initial creation of LD was very poorly estimated, which is the case with all LD mapping methods [16,21]. However, this does not affect the estimation of other parameters because t has little effect on the value of the likelihood function. The $\prod_Q(0)$ estimate is nearly the same for both methods. The large difference from the true value of $\prod_Q(0)$ is due to the simulation procedure that rejects the sample paths leading to the final frequency $\prod_Q(t)$ being smaller than 0.05. Using the Wright-Fisher model described in the Appendix, it can be proved that $\prod_Q(0)$ is equal to the expectation of $\prod_Q(t)$. Therefore, the empirical mean of $\prod_Q(0)$ over the 500 replicates is actually an estimate of the conditional expected value of $\prod_Q(t)$ given that $0.05 \le \prod_Q(t)$ and $\prod_Q(0) = 0.00125$. Using a diffusion approximation of the Wright-Fisher process and the corresponding probability density given in [32], we found that this quantity was equal to 0.105. The empirical mean of $\prod_Q(0)$ is in good agreement with this theoretical value, and the slight remaining bias might come from the selective advantage given to allele Q in the first few generations of our simulations, which is not accounted for in the diffusion approximation.

Table 2: Comparison of model parameter estimates.

Model parameter	True value	Empirical mean (standard error)	
		T2	HAPim
x (in cM)	3.6	3.73 (5.1e-2)	3.62 (4.3e-2)
$\prod_Q(0)$	0.00125	0.13 (3.4e-3)	0.12 (3.0e-3)
a	1	1.02 (2.5e-2)	0.98 (2.4e-2)
d	1	0.87 (3.2e-2)	0.97 (2.7e-2)
t	100	57.9 (8.5)	53.4 (4.6)
α	1	0.92 (6.3e-3)	0.92 (6.0e-3)

Empirical means (and their standard errors) of the model parameter estimates under the T2 and HAPim methods. The single nucleotide polymorphism (SNP) marker spacing was 0.25 cM, the effective population size was N = 400, and the initial association was complete.

Tables 3, 4, and 5 focus on a marker spacing of 0.125 cM, because the results of Table 1 indicate that the gain from using HAPim was greater with dense maps. We considered only biallelic markers, since in practice MSTs are rarely found with such a density. We investigated the role of (i) effective population size N (Table 3), and found that as N increases, the MSEs of both methods decrease but the difference between the methods becomes less significant; (ii) sample size (Table 4), and found that for N = 400 and N = 800, the gain of HAPim over T2 appears to recover since a sample from the population is used instead of the entire population; this gain was always significant, particularly with small samples; and (iii) time since the initial LD creation (Table 5), and found that when this time is small, the accuracy of both methods is limited; it is increased with larger evolution times, in which cases HAPim performed much better than T2; it is well known that short evolution times result in the high LD area extending to many markers around the QTL, which limits the accuracy of LD mapping methods in general.

Elucidating the mechanisms underlying the results of such simulations is extremely difficult, because parameters

Table 3: Effect of effective population size.

N	MSE		Difference in MSE P value
	T2	HAPim	
200	0.63	0.44	0.001**
400	0.52	0.44	0.135
800	0.30	0.29	0.782
1600	0.15	0.13	0.414

**: P < 0.01

Mean square errors (MSEs) in cM^2 of quantitative trait locus (QTL) location estimates obtained by the T2 and HAPim methods for various effective population sizes. The single nucleotide polymorphism (SNP) marker spacing was 0.125 cM, = 100 and the initial association was complete.

Table 5: Effect of time since initial creation of linkage disequilibrium (LD).

t	MSE		Difference in MSE P value
	T2	HAPim	
50	0.69	0.64	0.495
100	0.52	0.44	0.135
200	0.41	0.26	< 0.001**
300	0.25	0.17	0.005**

*: P < 0.05, ** : P < 0.01

Mean square errors (MSEs) in cM^2 of quantitative trait locus (QTL) location estimates obtained by the T2 and HAPim methods for various values of time *t* since initial LD creation. The single nucleotide polymorphism (SNP) marker spacing was 0.125 cM, *N* = 400, and the initial association was complete.

share complex interactions – increasing a particular parameter may have either a positive or a negative effect on the accuracy, depending on the value of the other parameters. Our model describes the decay in the LD from an initial event. In this context, we know that the accuracy of both LD methods mostly depends on the value of the product ct [3], with ct ≈ 2 being optimal. This may explain the results of Table 5. However, this explanation is only applicable to large values of N; for smaller values of N, at least two phenomena affect this rule. First, the approximation of the likelihood (3) is worse than with large N (but we do not know whether T2 or HAPim is affected the most). Second, the LD created by random drift along generations is no longer negligible, and its amount depends on the product Nc [29]. However, Tables 3 and 4 suggest that unless the sample size is very large (which also requires a very large effective population size), it is really worth using HAPim instead of T2. HAPim models the evolution of haplotype frequencies more precisely, which balances the lack of information.

Table 4: Effect of sample size.

N	N_s	MSE		Difference in MSE P value	Power	
		T2	HAPim		T2	HAPim
400	50	1.34	0.99	< 0.001**	0.36	0.59
	100	1.07	0.84	0.011*	0.66	0.88
	200	0.74	0.56	0.013*	0.88	0.99
800	100	1.08	0.87	0.033*	0.44	0.69
	200	0.66	0.49	0.008**	0.83	0.95
	400	0.42	0.31	0.006**	0.99	1.00

* : P < 0.05, ** : P < 0.01

Mean square errors (MSEs) in cM^2 of quantitative trait locus (QTL) location estimates and powers to detect the QTL obtained by the T2 and HAPim methods for various population and sample sizes. The single nucleotide polymorphism (SNP) marker spacing was 0.125 cM, *t* = 100, and the initial association was complete. The power was computed for a type I error of 0.05.

Table 4 also includes, for each effective population size and sample size, the power of HAPim and T2 to detect the QTL. This power was estimated from the same 500 replicates as the MSEs, using an approximate threshold as explained in the Methods section. As expected and observed in [33], the power was greater with greater sample size and with lower effective population size. The power results were also consistant with the MSE results: they revealed an important gain from using HAPim, that decreased as sample size increased. The number of replicates in which the log-likelihood ratio test was higher with HAPim than with T2 ranged from 80% to 90% depending on N and N_s. In Tables 3 and 5, this proportion was generally lower (even 50% with t = 300, Table 5) and the power obtained with both methods was always around 1. However the MSEs were still better with HAPim, which indicates that this method also allows a better discrimination between positions.

To complete our study, we compared the robustness of both methods to more complex evolution scenarios. In the first scenario, LD was initially created in a population in which allele *Q* already existed and was in linkage equilibrium with other markers. Since the degree of the initial association is strongly related to the number of alleles, we included both MST and SNP markers. We took a marker spacing of 0.25 cM and an effective population size N = 400, as previously done in Table 1. The results listed in Table 6 indicate that the MSEs were smaller than in the corresponding homogeneity scenario of Table 1, despite that heterogeneity decreased the strength of association between the QTL and marker alleles. This is probably due to the frequency of allele *Q* being higher in the heterogeneity scenario, which increases the percentage of the trait variance explained by the QTL and hence improves the

Table 6: Incomplete initial linkage disequilibrium (LD) scenario.

Marker type	MSE		Difference in MSE P value
	T2	HAPim	
Initial frequency of Q = 5%			
SNP(1)	0.99	0.68	0.031*
MST(2)	0.42	0.17	< 0.001**
Initial frequency of Q = 10%			
SNP(1)	0.95	0.64	0.039*
MST(2)	0.63	0.20	< 0.001**

* : P < 0.05, ** : P < 0.01
(1) : single nucleotide polymorphism
(2) : microsatellite
Mean square errors (MSEs) in cM^2 of quantitative trait locus (QTL) location estimates obtained by the T2 and HAPim methods for various heterogeneity parameter values, t = 100, N = 400, and a marker spacing of 0.25 cM.

mapping precision. HAPim strongly outperformed T2, particularly for MSTs.

In the second scenario we introduced phenocopies. As in the heterogeneity scenario, we chose N = 400, a marker spacing of 0.25 cM, and both SNP and MST markers. The MSEs with this scenario, given in Table 7, were much larger than in the corresponding scenario of Table 1, particularly for SNPs. MSTs are less affected by phenocopies because the number of possible marker haplotypes that can be carried by a "false Q" individual is much larger than with SNPs. The risk of the method producing a false-positive error is thus reduced. Using HAPim instead of T2 also reduces this risk, because the allele frequencies at flanking markers are modeled jointly. In this scenario, HAPim clearly outperformed T2.

Table 7: Scenario with phenocopies.

Marker type	MSE		Difference in MSE P value
	T2	HAPim	
Phenocopy rate = 15% [b]			
SNP(1)	2.65	2.03	0.021*
MST(2)	0.84	0.35	< 0.001**
Phenocopy rate = 30%			
SNP(1)	4.90	3.29	< 0.001**
MST(2)	1.94	0.67	< 0.001**

* : P < 0.05, ** : P < 0.01
[b]: Phenocopy rate refers to the percentage of q alleles in the last generation that have given the same phenotype as the Q allele
(1) : single nucleotide polymorphism
(2) : microsatellite
Mean square errors (MSEs) in cM^2 of quantitative trait locus (QTL) location estimates obtained by the T2 and HAPim methods for various phenocopy rate values, t = 100, N = 400, and a marker spacing of 0.25 cM.

Comparison with other haplotype methods

Modeling the information from haplotypes consisting of more than two markers may improve the precision of location estimates. Therefore, further simulations were carried out to compare our HAPim method with the IBD method of Meuwissen and Goddard [21]. Their method is one of the most classical full-haplotype methods, and the similarity of their genetic model to ours makes the comparison easier than for coalescent-based methods such as in [20,23]. We duplicated the simulation scenarios described in Table 2 in [21]: 50 population replicates with biallelic markers initially at equal frequencies with spacings of 0.25, 0.5, and 1.0 cM, an effective population size and a sample size of $N = N_s = 100$, and a time t = 100 since the initial mutation. The QTL was in the middle of the chromosome region. In order for the results to be perfectly comparable, the mutant allele was not given a slight selective advantage after the mutation time (in contrast to previous simulation scenarios, as explained in the Methods section). Table 8 presents the distribution of the deviations (in marker intervals) in the QTL location estimates from the correct bracket. The results can be directly compared with those of Table 3 in [21]. A chi-square test of equality between the deviation distributions of HAPim and [21] revealed no significant difference (the smallest p

Table 8: Comparison with the IBD method of Meuwissen and Goddart.

Marker spacing (cM)	Deviation				
	0	1	2	3	4
frequency of allele Q ≥ 0.1					
1.0	16	17	9	5	3
0.5	12	20	10	2	6
0.25	12	18	8	6	6
frequency of allele Q ≥ 0					
1.0	15	14	6	8	7
0.5	10	17	12	7	4
0.25	11	14	11	5	9

Distribution of the deviations (in marker brackets) of the quantitative trait locus (QTL) location estimates from the correct bracket for the HAPim method under the default simulation scenarios (biallelic markers with N = 100, N_s = 100, and t = 100) described in [1]. A deviation of 0 means the estimated position was in the correct marker bracket, 1 means the estimated position was one bracket away from the correct position, etc.

BMC Genomics 2006, **7**:54 http://www.biomedcentral.com/1471-2164/7/54

value was 0.08), and a t-test on the MSEs of both methods also did not reveal any significant difference.

We also tested our method under the simulation scenarios used by Grapes and colleagues [26,27], who compared single- and two-marker regression analysis with an IBD method very similar to that in [21]. For the same number of markers, the least-square mean absolute differences (LSMDs) between the estimated and the true QTL location were clearly smaller with the IBD full-haplotype method ([26], Table 2), which confirms its superiority. A subsequent study [27] revealed that mapping precision of the IBD method could be increased by using a smaller window of markers (four or six), and that using a window of only two markers provided the same accuracy as using the full haplotype (ten markers). We reproduced these simulation scenarios using the same number of replicates (1000) as they used. The results we obtained with HAPim were similar to the ones given by their IBD method using two-marker haplotypes: LSMDs of 1.36, 0.71, and 0.39 for marker spacings of 1.0, 0.5, and 0.25 cM, respectively.

Discussion

The present simulation study focused on particular values of model parameters, and hence the revealed good properties of HAPim may not hold for other values. However, we consider that the range of parameter values explored includes most of the situations where LD information can be used efficiently for mapping. For instance, the largest value of t we considered was 300 (Table 5), and whilst many favorable mutations are much older than 300 generations, it is very unlikely for a population to satisfy the strong hypotheses of the assumed Wright-Fisher model (e.g., random mating and no migrations) over such a long period. In many cases a strong founder effect occurred quite recently, and this event then corresponds to time 0 in our method. In other situations, we know that recurrent mutations or migrations have occurred continuously in the population and consequently perturbed the LD structure. It is very likely that no method could exploit the LD information for mapping in such cases [30,31].

We consider effective population sizes between 100 and 1600 to be realistic for most breeding populations, where the high level of inbreeding reduces the effective size. The effective size of the isolated human populations typically used in LD studies (e.g., Finnish or Caucasian) is generally around 10,000 [10]. We were not able to study such cases, but extrapolating the results of Table 3 leads to the supposition that there is no difference between T2 and HAPim for such large populations, provided that the marker spacing remains larger than around 0.1 cM. Another specific feature of such isolated human populations is their exponential growth rate. It would be easy to include this in our model, but it would have no effect as long as the first-order approximation of the likelihood (4) is used [10]. Another case that we did not study is that of very dense maps (marker spacing smaller than 0.01 cM). In that case the flanking marker haplotypes probably lose relevant information contained in full haplotypes, and modeling the information from more than two markers may improve the mapping precision. Our method could be extended by replacing – on each side of the QTL – the flanking marker by a flanking haplotype, and then performing the computations exactly as before. The extension is straightforward if we assume linkage equilibrium between all markers, but an increased precision is not guaranteed since background LD is not accounted for. As an alternative to assuming equilibrium, one could model marker allele frequencies along the chromosome as a first-order Markov chain with parameters estimated from the marker data at time t [12,13], but it would be more difficult to integrate this change in the derivations given in the Appendix.

The model itself and its hypotheses can be criticized. For example, we assume that the marginal allele frequencies are constant and that markers are in linkage equilibrium; i.e., $\Pi_{i_1,i_2} = \Pi_{i_1}\Pi_{i_2}$. Actually, the expression we obtained for $\mathbb{E}[\Pi_{i_1,Q,i_2}(t)]$ would be the same if we only assumed equilibrium at time 0 between markers. Considering only the first moment of haplotype frequencies, as we do in (4), this is the best we can do. Accounting for the LD between markers would thus require consideration of a second-order approximation of the likelihood and of the variances of the haplotype frequencies. This may improve the performance of the method, whereas no improvement was observed in [10]. In our simulations the marker frequencies were not constant and the equilibrium imposed at the first simulated generation was randomly broken by drift in the few generations until the time of the mutation. Thus, at time 0 the markers were not in equilibrium. One other strong approximation of the model is the absence of mutations or selection. While the effect of mutations is often negligible on the short evolution times we are interested in, they could be easily accounted for in the derivations of $\mathbb{E}[\Pi_{i_1,Q,i_2}(t)]$ using a stepwise mutation model [34]. Selection advantages for Q or q would be more difficult to incorporate, because they make the expression of $\mathbb{E}[\Pi_{i_1,Q,i_2}(t+1)]$ (see (8) in Appendix) non linear in $\Pi(t)$. Finally, it should be noted that $\Pi_Q(t)$ was not assumed to be constant in our model; this assumption was made in [10] and criticized in [35].

Knowledge of the haplotypes is required to apply haplotype-based mapping methods including the one described here. In our simulations we used a true set of haplotypes, but in the analysis of real data the haplotypes have to be inferred from the data or using pedigree information. Several algorithms have been proposed in the literature to perform such inferences [36]. Combined advances in both these algorithms and molecular haplotyping methods will enable this question to be solved more efficiently in the future. Moreover, several studies [37,38] have shown that the efficiency of fine mapping methods is not greatly reduced by uncertainty of the haplotype phases. If this did not hold for HAPim, the gain from using this method rather than T2 would be low given that T2 is not affected by the haplotype phases. This should be investigated in the future.

In our simulation study the results obtained with HAPim were similar to the ones given by the IBD method using two-marker haplotypes. Nevertheless, there are fundamental differences between HAPim and the IBD method. First, haplotype effects are modeled as fixed effects in the former and as random effects in the latter. While it is well-known that location parameters are easier to estimate than dispersion parameters, it is not clear whether this has a significant effect on the estimation of the QTL position. Second, the IBD method doesn't include dominance effects, while HAPim handles that very efficiently, as illustrated in Table 2. Third, the time t since the initial creation of LD and the effective population size N have to be known before using the IBD method. Some simulation results in [21] suggested that the default choice of $t = 100$ and $N = 100$ was almost optimal, whatever the true value of these parameters. However the comparison of tables V and VII in [22] indicates that the IBD matrix with $N = 1000$ is really different from the one with $N = 100$. Thus it is not obvious why the IBD method assuming $N = 100$ should be accurate for a population of actual effective size $N = 1000$. On the other hand, neither t nor N are required for the use of HAPim. Consequently this method can be used in a wider range of populations. A nice advantage of the IBD method is its ability to deal with haplotypes composed of more than two markers. If used with caution, this can provide more accuracy in location estimates [27]. As explained previously, HAPim could also offer this possibility in the future. At present, the several differences highlighted in this paragraph already justify the interest of this method.

An important purpose of QTL mapping methods is to provide a confidence interval for the QTL location. Classical pedigree linkage analyses have proposed log-odds (LOD)-support intervals [39], similar confidence intervals [40], and bootstrap confidence intervals [41]. The simplicity of the bootstrap technique, its ease of implementation, and the accuracy of the coverage probability makes it an appealing approach to use. In LD mapping methods, the coverage accuracy of the LOD-support interval and the credible interval in the Bayesian framework have been studied only for disease traits [12,23,42]. Simulations have shown that both intervals are either unbiased or only slightly conservative. This issue has not yet been addressed for QTL location. An anticonservative bootstrap confidence interval was obtained when we ran a preliminary single simulation with HAPim, which may indicate that the classical bootstrap scheme we used – sampling with replacement of entire records – did not produce enough variability of the QTL location estimate. Confirmation of this result may indicate that providing a correct confidence interval for the QTL location is a challenging and tricky problem.

Although our two-marker haplotype model was basically designed for unrelated individuals, it can also be used in situations where pedigree information is available. For instance, in studies involving large half-sib families, our model can easily be integrated in the combined LD and linkage mapping method of Farnir and colleagues [19]. In their method, LD information is contained in the probabilities of Table 1 ([19], p. 277). These probabilities were derived under a single-marker model, and could instead be derived under our two-marker model using (5) without changing the rest of the method. However, the use of combined LD and pedigree information appears to be more efficient in designs with many small families than in those with a few large families [43]. Consequently a promising strategy for future QTL mapping studies would be to genotype and phenotype more unrelated individuals and use the parental information (if any is available) to infer the haplotypes. In this context the use of our method could be fruitful.

Conclusion

We have presented a new method for the fine mapping of QTLs, denoted HAPim. It is a likelihood method, whose originality is in modeling the frequencies of haplotypes comprising one trait locus and two flanking markers. Theoretical derivations under this evolution model avoid the intensive computations required to evaluate the likelihood values at each location.

Our simulations have demonstrated the excellent properties of our method. Over a wide range of parameter values (effective population sizes and sample sizes from 200 to 1600, times since LD creation from 50 to 300 generations, and marker spacings from 0.125 to 2 cM), the MSEs obtained with HAPim were almost always significantly lower than those obtained with composite likelihood method T2. Combined with a previous study [18], these results show that HAPim is more accurate than single-

BMC Genomics 2006, **7**:54 http://www.biomedcentral.com/1471-2164/7/54

marker methods and composite likelihood methods in general. The power to detect the QTL was also greater with HAPim. With approximately the same parameter values, we observed that HAPim was as accurate as the classical IBD method [21] used with two- or ten-marker haplotypes. It also has several advantages over the IBD method, as the ability to incorporate dominance effects and to deal as easily with any value of t or N. Finally, our simulations suggested that the use of MSTs is very efficient if the analysis is performed with HAPim: the computing time was longer than with SNPs but was still reasonable, and the estimates were more robust to departures from the assumed model. Given that more and more mapping studies are being designed with SNP, this suggests that close SNPs should be combined into groups of two or three to build pseudo-multiallelic markers that avoid spurious associations.

Our method could be improved in several ways, such as by modeling mutations or LD between markers, and using haplotypes with more than two markers, but it is unclear whether these modifications would increase the precision. Providing confidence intervals – in addition to the pointwise QTL location estimates – will also be an interesting challenge. The continuing advances in genotyping and haplotyping technologies will increase the importance of LD fine mapping methods, even in situations where pedigree information is available.

Methods

Likelihood maximization

The description of the model highlights that parameters other than the QTL location x have to be estimated: the time t since the initial creation of LD, the initial frequency $\prod_Q(0)$ of allele Q, the initial associated haplotype j, and the heterogeneity parameter α. We take the values that satisfy

$$\max_{x,t,\Pi_Q(0),j,\alpha} \mathcal{L}(x,t,\Pi_Q(0),j,\alpha \mid)$$

This maximization is carried out numerically using the E04CCF simplex algorithm from the NAG library [44]. Marker allele frequencies Π_{i_1} and Π_{i_2} also have to be estimated. We use their empirical frequencies in the sample and thus do not need to include them in the likelihood maximization.

We also tested a homogeneity method where a was arbitrarily set to 1. On the basis of simulation results (similar to those presented in this paper), we finally dropped this because it was not as robust as the more general method to departures from the assumed model.

Simulation procedure

We used forward simulations as outlined in [18,45]. The baseline scenario was as follows. We initially define a population of $2N$ haplotypes with L equally spaced markers, either biallelic (SNPs) or multiallelic (MSTs) with five alleles. In both cases, all of the marker alleles have the same frequency and the markers are in linkage equilibrium. Then, each new generation is created by sampling N pairs of haplotypes at random from the current generation and allowing random recombinations within these pairs. The recombination rate for each marker interval is computed using Haldane's mapping function. We let the population evolve for $20 \times (N/400)$ generations in order to break the linkage equilibrium between markers with a random drift force that does not depend on the effective population size. At time 0, a mutated allele Q is introduced at the QTL location on a single haplotype, and again we let the population evolve as previously. At time t, a sample of N_s individuals is collected, and phenotypes for the trait are simulated according to the model in (1), with $a = 1$, $d = 1$ (complete dominance) and $\sigma^2 = 1$. In all simulation scenarios but the one reported in Table 3, the sample size N_s was equal to the effective population size N.

Two extensions of this scenario were also considered. Firstly, some copies of allele Q were introduced into the population from the first generation of the simulation, with frequency $\prod_Q(0)$ equal to 0.05 or 0.10. These earlier copies of Q were in equilibrium with all markers, so at time 0 the association created between Q and one particular marker haplotype was incomplete. Secondly, we allowed the presence of phenocopies; i.e., phenotypes that mimic the phenotype produced by the mutation. To reproduce this effect, a given percentage of the individuals carrying allele q (15% or 30%) were randomly drawn in the last generation and were given the same genetic effect as individuals carrying allele Q.

In all scenarios, replicates were discarded when fixation occurred for the QTL or any of the markers, or when the final frequency of allele Q was less than 0.05 or greater than because rare QTL alleles account for a small proportion of the trait variance and are not of interest in QTL mapping studies. To reduce the number of discarded replicates, the new QTL allele was conferred with a slight selective advantage during a few generations after time 0.

The accuracy of QTL location estimates was evaluated according to the MSE defined as

$$MSE = \frac{1}{R}\sum_{r=1}^{R}(x_r - x)^2$$

where R is the number of replicates (equal to 500 unless otherwise specified), $\hat{x}_r$ is the estimated QTL location in the rth replicate, and x is the true location. The MSE contains information of both the bias and the variance of location estimates. Differences in MSE between methods were tested using paired t-tests while assuming normality.

Power computation

Together with the set of optimal parameter values, HAPim returns the log-likelihood ratio test between the null hypothesis "$a = d = 0$" and its alternative. In order to compare the power of T2 and HAPim we computed an approximate threshold for any set of population parameter values (N, t, marker spacing …). This threshold was obtained as the empirical 0.95 quantile of 500 replicates under the null hypothesis.

Authors' contributions

SB and JA contributed equally to this work. SB developed the mathematical description and JA wrote the computer programs, and they both were involved in the preparation of the draft manuscript. All authors participated in the design conception, the interpretation of the simulation results, and the elaboration of the manuscript under the leadership of BM.

Appendix

Derivation of the formula for $\mathbb{E}[\Pi_{i,Q}(t)]$

In this section we consider the segregation of one QTL and one multiallelic marker, with a recombination rate c between them. Let $X_{i,Q}(t)$ and $X_{i,q}(t)$ be the number of haplotypes (i, Q) and (i, q) in the population at generation t, respectively, and $\mathbf{X}(t) = (X_{1,Q}(t),\ldots, X_{I,Q}(t), X_{1,q}(t),\ldots, X_{I,q}(t))$; we define also the vector of haplotype frequencies

$$\Pi(t) = \frac{X(t)}{2N(t)} = (\Pi_{1,Q}(t),\cdots;\Pi_{I,Q}(t),\Pi_{1,q}(t), \quad ,\Pi_{I,q}(t))$$

These vectors are stochastic processes of time. We first present a two-locus Wright-Fisher model [46,47] that describes the distribution of $\mathbf{X}(t + 1)$ given $\mathbf{X}(t)$. From this model and under the assumption that the allelic frequency $\prod_i(t) = \prod_{i,Q}(t) + \prod_{i,q}(t)$ is deterministic and time invariant, we deduce a recursive relation between $\mathbb{E}[\prod(t + 1)]$ and $\mathbb{E}[\prod(t)]$ that we use to determine the expression for $\mathbb{E}[\prod_{i,Q}(t)]$.

In the two-locus Wright-Fisher model, the effective population size $N(t)$ is a deterministic function of time and the vector $\mathbf{X}(t + 1)$ follows, conditional on $\mathbf{X}(t)$, a multinomial distribution with parameters $(2N(t + 1), r_{1,Q}(t),\ldots, r_{I,Q}(t), r_{1,q}(t),\ldots, r_{I,q}(t))$, where

$$r_{i,Q}(t) = (1 - c)\prod_{i,Q}(t) + c\prod_Q(t)\prod_i(t)$$

The two terms of this formula represent the probabilities of choosing nonrecombining and recombining haplotypes.

From the properties of multinomial distributions we have $\mathbb{E}[X_{i,Q}(t + 1) \mid \mathbf{X}(t)] = 2N(t + 1)r_{i,Q}(t)$, and thus $\mathbb{E}[\prod_{i,Q}(t + 1) \mid \mathbf{X}(t)] = r_{i,Q}(t)$. A classical result on conditional probabilities yields

$$\begin{aligned}\mathbb{E}[\Pi_{i,Q}(t+1)] &= \mathbb{E}[\mathbb{E}[\Pi_{i,Q}(t+1) \mid \mathbf{X}(t)]] \\ &= \mathbb{E}[r_{i,Q}(t)] \\ &= (1-c)\mathbb{E}[\Pi_{i,Q}(t)] + c\mathbb{E}[\Pi_Q(t)\Pi_i(t)]\end{aligned}$$

We assume that $\prod_i(t) = \prod_i$ is time invariant, which is reasonable because allele i is supposed to be much older than allele Q and consequently its frequency is much higher. This leads to

$$\mathbb{E}[\prod_{i,Q}(t + 1)] = (1 - c)\,\mathbb{E}[\prod_{i,Q}(t)] + c\,\mathbb{E}[\prod_Q(t)]\prod_i$$

and the entire vector $\prod(t)$ satisfies

$$\mathbb{E}[\prod(t + 1)] = \mathbb{E}[\prod(t)]\,(cA + (1 - c)Id_I) \quad (6)$$

where $A = (\prod_1,\ldots, \prod_I) \otimes \mathbb{1}_I$, where $\otimes$ is the Kronecker product, $\mathbb{1}_I$ is the column vector of size I with all components equal to 1, and Id_I is the identity matrix of size $I \times I$.

A is idempotent since $\sum_{i=1}^{I}\prod_i = 1$, and so we can prove by recurrence on t that

$$\mathbb{E}[\prod(t)] = \mathbb{E}[\prod(0)]\,((1 - (1 - c)^t)\,A + (1 - c)^t\,Id_I)$$

Taking the ith coordinate we get

$$\mathbb{E}[\prod_{i,Q}(t)] = (1 - c)^t\prod_{i,Q}(0) + (1 - (1 - c)^t)\,\prod_Q(0)\prod_i \quad (7)$$

Derivation of the formula for $\mathbb{E}[\Pi_{i_1,Q,i_2}(t)]$

We now consider the more complex case of two multiallelic markers flanking the QTL. We proceed as in the previous section, defining first a three-locus Wright-Fisher model and then deducing from it a recurrence relation for the expected value of haplotype frequencies. To do this we

BMC Genomics 2006, **7**:54 http://www.biomedcentral.com/1471-2164/7/54

also assume that the markers are in equilibrium. From the recurrence relation we finally obtain the expression for $\mathbb{E}[\Pi_{i_1,Q,i_2}(t)]$.

The three-locus Wright-Fisher model describes the segregation of haplotypes composed of the QTL and two flanking markers. The first marker has I_1 alleles and a recombination rate c_1 with the QTL; the second one has I_2 alleles and a recombination rate c_2 with the QTL. We denote $X_{i_1,Q,i_2}(t)$, $i_1 = 1,\ldots, I_1$, $i_2 = 1,\ldots, I_2$, as the number of copies of haplotype (i_1, Q, i_2) in the population at generation t, and $\Pi_{i_2,Q,i_2}(t)$ as the corresponding frequency. $\mathbf{X}(t + 1)$ has dimension $2I_1I_2$, but still has a multinomial distribution given $\mathbf{X}(t)$ with parameters $(2N(t+1), r_{1,Q,1}(t), \cdots; r_{I_1,Q,I_2}(t), r_{1,q,1}(t), \quad , r_{I_1,q,I_2}(t))$, where

$r_{i_1,Q,i_2}(t) = (1-c_1)(1-c_2)\Pi_{i_1,Q,i_2}(t) + c_1(1-c_2)\Pi_{i_1}\Pi_{Q,i_2}(t) + c_2(1-c_1)\Pi_{i_2}\Pi_{i_1,Q}(t) + c_1c_2\Pi_{i_1,i_2}(t)\Pi_Q(t)$

, Π_{i_1} , Π_{i_2} , and $\prod_Q(t)$ are the marginal frequencies of alleles i_1 at the left marker, i_2 at the right marker, and Q at the QTL, respectively, and $\Pi_{i_1,Q}(t)$ and $\Pi_{Q,i_2}(t)$ are the marginal frequencies of haplotypes (i_1, Q) and (Q, i_2), respectively. The four terms in this formula correspond to the different origins of haplotypes $(i_1 Q, i_2)$ at generation t + 1: nonrecombining, recombining between QTL and the left-side marker, recombining between QTL and the right-side marker, and double recombining.

As in the previous section, we can express the expected value of the frequencies of haplotypes at time t + 1 as

$$\begin{aligned}\mathbb{E}[\Pi_{i_1,Q,i_2}(t+1)] &= \mathbb{E}[\,\mathbb{E}[\Pi_{i_1,Q,i_2}(t+1) \mid X(t)]]\\ &= \mathbb{E}[r_{i_1,Q,i_2}(t)]\\ &= (1-c_1)(1-c_2)\mathbb{E}[\Pi_{i_1,Q,i_2}(t)] + c_1(1-c_2)\Pi_{i_1}\,\mathbb{E}[\Pi_{Q,i_2}(t)]\\ &+ c_2(1-c_1)\Pi_{i_2}\mathbb{E}[\Pi_{i_1,Q}(t)] + c_1c_2\,\mathbb{E}[\Pi_{i_1,i_2}(t)\Pi_Q(t)]\end{aligned}$$

Assuming that the markers are in equilibrium and that the allelic frequencies are constant; i.e.,

$$\Pi_{i_1,i_2}(t) = \Pi_{i_1}(t)\Pi_{i_2}(t) = \Pi_{i_1}\Pi_{i_2}$$

we get

$$\begin{aligned}\mathbb{E}[\Pi_{i_1,Q,i_2}(t+1)] &= (1-c_1)(1-c_2)\,\mathbb{E}[\Pi_{i_1,Q,i_2}(t)] + c_1(1-c_2)\,\mathbb{E}[\Pi_{Q,i_2}(t)]\Pi_{i_1}\\ &+ c_2(1-c_1)\mathbb{E}[\Pi_{i_1,Q}(t)]\Pi_{i_2} + c_1c_2\,\mathbb{E}[\Pi_Q(t)]\Pi_{i_1}\Pi_{i_2}\end{aligned}$$

Substituting $\mathbb{E}[\Pi_{i_1,Q}(t)]$ and $\mathbb{E}[\Pi_{Q,i_2}(t)]$ with the expressions determined in the previous section gives

$$\begin{aligned}\mathbb{E}[\Pi_{i_1,Q,i_2}(t+1)] &= (1-c_1)(1-c_2)\,\mathbb{E}[\Pi_{i_1,Q,i_2}(t)] + \beta_2 c_1(1-c_2)^{t+1} + \beta_1 c_2(1-c_1)^{t+1}\\ &+ (c_1(1-c_2) + c_2(1-c_1) + c_1c_2)\Pi_Q(0)\Pi_{i_1}\Pi_{i_2}\end{aligned} \quad (8)$$

This is a recurrence relationship that can be solved easily. We can prove that if $(u_t)_{t\geq 0}$ is a series in $\mathbb{R}$ defined by

$$u_{t+1} = au_t + b\alpha^{t+1} + c\gamma^{t+1} + d$$

then for every $t \geq 0$,

$$u_t = a^t u_0 + b\sum_{s=1}^{t} a^{t-s}\alpha^s\gamma + c\sum_{s=1}^{t} a^{t-s}\quad^s + d\frac{1-a^t}{1-a}$$

Applying this result with $a = (1 - c_1)(1 - c_2)$, $b = \beta_2 c_1$, $\alpha = 1 - c_2$, $c = \beta_1 c_2$, $\gamma = 1 - c_1$, and $d = (c_1 + c_2 - c_1c_2)\prod_Q(0)\ \Pi_{i_1}\ \Pi_{i_2}$ yields

$$\begin{aligned}\mathbb{E}[\Pi_{i_1,Q,i_2}(t)] &= (1-c_1)^t(1-c_2)^t\Pi_{i_1,Q,i_2}(0)\\ &+ \beta_2 c_1(1-c_2)^t\left(\sum_{s=1}^{t}(1-c_1)^{t-s}\right) + \beta_1 c_2(1-c_1)^t\left(\sum_{s=1}^{t}(1-c_2)^{t-s}\right)\\ &+ (c_1 + c_2 - c_1c_2)\Pi_Q(0)\frac{1-(1-c_1)^t(1-c_2)^t}{1-(1-c_1)(1-c_2)}\Pi_{i_1}\Pi_{i_2}\\ &= (1-c_1)^t(1-c_2)^t\Pi_{i_1,Q,i_2}(0)\\ &+ \beta_2\beta(1-c_2)^t\left(1-(1-c_1)^t\right) + {}_1(1-c_1)^t\left(1-(1-c_2)^t\right)\\ &+ \Pi_Q(0)(1-(1-c_1)^t(1-c_2)^t)\Pi_{i_1}\Pi_{i_2}\end{aligned}$$

Replacing β_1 and β_2 by their actual expressions gives

$$\begin{aligned}\mathbb{E}[\Pi_{i_1,Q,i_2}(t)] &= \Pi_Q(0)\Pi_{i_1}\Pi_{i_2} + (1-c_1)^t(\Pi_{i_1,Q}(0) - \Pi_Q(0)\Pi_{i_1})\Pi_{i_2} + (1-c_2)^t(\Pi_{Q,i_2}(0) - \Pi_Q(0)\Pi_{i_2})\Pi_{i_1}\\ &+ (1-c_1)^t(1-c_2)^t(\Pi_{i_1,Q,i_2}(0) - \Pi_{k,Q}(0)\Pi_{i_2} - \Pi_{Q,i_2}(0)\Pi_{i_1} + \Pi_Q(0)\Pi_{i_1}\Pi_{i_2})\end{aligned} \quad (9)$$

Acknowledgements

This work was partially funded by the French Ministry of Research (Ministère de la Recherche) under the project Bioinformatique awarded on June 2000.

References

1. Bodrner W: **Human genetics: the molecular challenge.** *Cold Sring Harbor Symp Quant Biol* 1986, **51:**1-13.
2. Boehnke M: **Limits of resolution of genetic linkage studies: implication for the positional cloning of human disease genes .** *Am J Hum Genet* 1994, **55:**379-390.
3. Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E: **Linkage disequilibrium mapping in isolated founder populations: diastrophic dysphasia in Finland.** *Nat Genet* 1992, **2:** 204-211.
4. Jorde L: **Linkage disequilibrium as a gene-mapping tool.** *Am J Hum Genet* 1995, **52:**11-14.
5. Cox T, Kerem B, Rommens J, Iannuzzi M, Drumm M, Collins F, Dean M, *et al.*: **Mapping of the cystic fibrosis gene using putative ancestral recombinants.** *Am J Hum Genet* 1989:A136.
6. Theilman J, Kanani S, Shiang R, Robbins C, Quarrell O, Huggins M, Hedrick A, Weber B, Collins C, Wasmuth J: **Non-random associa-**

tion between alleles detected at D4S95 D4S98 and the Huntington's disease gene. *J Med Genet* 1989, **26:**676-681.
7. MacDonald M, Novelletto A, Lin C, Tagle D, Barnes G, Bates G, Taylor S, Allitto B, Altherr M, Myers R, Lehrach H, Collins F, Wasmuth J, Frontali M, Gusella J: **The Huntington's disease candidate region exhibits many different haplotypesgene.** *Nat Genet* 1992, **1:**99-103.
8. Kaplan N, Hill W, Weir B: **Likehood methods for locating disease genes in nonequilibrium populations.** *Am J Hum Genet* 1995, **56:**18-32.
9. Terwilliger J: **A powerfull likehood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci.** *Am J Hum Genet* 1995, **56:**777-787.
10. Xiong M, Guo S: **Fine scale genetic mapping based on linkage disequilibrium: theory and applications.** *Am J Hum Genet* 1997, **60:**1513-1531.
11. Collins A, Morton N: **Mapping a disease locus by allelic association.** *Proc Natl Acad Sci USA* 1998, **95:**1741-1745.
12. McPeak M, Strahs A: **Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine scale genetic mapping.** *Am J Hum Genet* 1999, **65:**858-875.
13. Morris A, Whittaker J, Balding D: **Bayesian fine-scale mapping of disease loci by hidden Markov models.** *Am J Hum Genet* 2000, **67:**155-169.
14. Graham J, Thompson E: **Disequilibrium likelihoods for fine-scale mapping of a rare allele.** *Am J Hum Genet* 1998, **63:**1517-1530.
15. Rannala B, Reeve J: **High resolution multipoint linkage disequilibrium mapping in the context of a human genome sequence.** *Am J Hum Genet* 2001, **69:**159-178.
16. Morris A, Whittaker J, Balding D: **Fine-scale mapping of disaese loci via shattered coalescent modelling of genealogies.** *Am J Hum Genet* 2002, **76:**686-707.
17. Boerwinkle E, Chakraborty R, Sing C: **The use of measured phenotype information in the analysis of quantitative phenotypes in man.** *Ann Hum Genet* 1986, **50:**181-194.
18. Abdallah J, Mangin B, Goffinet B, Cierco-Ayrolles C, Pérez-Enciso M: **A comparison between methods for linkage disequilibrium fine mapping of quantitative trait loci.** *Genet Res* 2004, **83:** 41-47.
19. Farnir F, Grisart B, Coppieters W, Riquet J, Berzi P, Cambisano N, Karim L, Mni M, Moisio S, Simon P, Wagenaar D, Vilkki J, Georges M: **Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk productionon bovine chromosome 14.** *Genetics* 2002, **161:**275-287.
20. Pérez-Enciso M: **Fine mapping of complex trait genes combinig pedigree and linkage disequilibrium information: a bayesian unified framework.** *Genetics* 2003, **163:**1497-1510.
21. Meuwissen T, Goddard M: **Fine mapping of quantitative trait loci using linkage disequilibrium with closely linked marker loci.** *Genetics* 2000, **155:**421-430.
22. Meuwissen T, Goddard M: **Prediction of identity by descent probabilities from marker-haplotypes.** *Genet Sel Evol* 2001, **33:** 605-634.
23. Zöllner S, Pritchard J: **Coalescent-based association mapping and fine mapping of complex trait loci.** *Genetics* 2005, **169:** 1071-1092.
24. Nordborg M: **Coalescent theory.** In *Handbook of statistical genetics* Edited by: Balding D, Bishop M, Cannings C. Wiley; 2001:179-212.
25. Blott S, Kim J, Moisio S, Schmidt-Kiintzel A, Cornet A, Berzi P, Cambisano N, Ford C, Grisart B, Johnson D, Karim L, Simon P, Snell R, Spelman R, Wong J, Vikki J, Georges M, Farnir F, Coppieters W: **Molecular dissection of a quantitative trait locus : a phenylalaline-to-tyrosine substitution in the transmembrane domain of the bovine groth hormone receptoris associated with a major effect on milk yield and composition.** *Genetics* 2003, **163:**253-266.
26. Grapes L, Dekkers J, Rothschild M, Fernando R: **Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci.** *Genetics* 2004, **166:**1561-1570.
27. Grapes L, Firat M, Dekkers J, Rothschild M, Fernando R: **Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity-by-descent.** *Genetics* in press.
28. Falconer D, Mackay T: *Introduction to quantitative genetics. Longman* 4 1996.
29. Hill W, Weir B: **Maximum-likelihood estimation of gene location by linkage disequilibrium.** *Am J Hum Genet* 1994, **54:** 705-714.
30. Kruglyak L: **Prospects for whole-genome linkage disequilibrium mapping of common disease genes.** *Nat Genet* 1999, **22:** 139-144.
31. Baret P, Hill W: **Gametic disequilibrium mapping: potential applications in livestock.** *Animal Breeding abstracts* 1997, **65:** 309-318.
32. Kirnura M: **Solution of a process of random genetic drift with a continuous model.** *Proc Nat Acad Sci USA* 1955, **41:**144-150.
33. Long A, Langley C: **The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits.** *Genome Res* 1999, **9:**720-731.
34. Ethier S, Kurtz T: *Markov processes. Characterization and convergence* Wiley series in probability and mathematical statistics, Wiley and Sons, Inc; 1986.
35. Rannala B, Slatkin M: **Likelihood analysis of disequilibrium mapping, and related problems.** *Am J Hum Genet* 1998, **62:**459-473.
36. Niu T: **Algorithms for inferring haplotypes.** *Genetic Epidemiology* 2004, **27:**334-347.
37. Morris A, Whittaker J, Balding D: **Little loss information due to unknown phase for fine-scale linkage disequilibrium mapping with single-nucleotide-polymorphism genotype data.** *Am J Hum Genet* 2004, **74:**945-953.
38. Lee S, van der Werf J: **The role of pedigree information in combined linkage disequilibrium and linkage mapping of quantitative trait loci in a general complex pedigree.** *Genetics* 2005, **169:**455-466.
39. Lander B, Botstein D: **Mapping mendelian factors underlying quantitaive traits using RFLP linkage maps.** *Genetics* 1989, **121:**185-199.
40. Mangin B, Goffinet B, Rebaï A: **Constructing confidence intervals for QTL location.** *Genetics* 1994, **138:**1301-1308.
41. Visscher P, Thompson R, Haley C: **Confidence intervals in QTL mapping by bootstrapping.** *Genetics* 1996, **143:**1013-1020.
42. Lam J, Roeder K, Devlin B: **Haplotype fine mapping by evolutionary trees.** *Am J Hum Genet* 2000, **66:**659-667.
43. Lee S, Julius H, van der Werf J: **The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage.** *Genet Sel Evol* 2004, **36:**145-161.
44. Group NA: *The NAG-Fortran library manual-mark 19* NAG Ltd; 1990.
45. Abdallah J, Goffinet B, Cierco-Ayrolles C, Pérez-Enciso M: **Linkage disequilibrium fine mapping of quantitative trait loci. A simulation study.** *Genet Sel Evol* 2003, **35:**513-532.
46. Karlin S, McGregor J: **Rates and probabilities of fixation for two locus random mating finite populations without selection.** *Genetics* 1968, **58:**141-159.
47. Ethier S, Nagylaki T: **Diffusion Approximations of the two-locus Wright-Fisher model.** *J Math Biol* 1989, **27:**17-28.

Chapitre 9

Compléments et perspectives

L'un des points centraux de la méthode d'estimation présentée au chapitre 8 est l'expression que nous avons obtenue pour $\mathbb{E}[\Pi_{i_1,Q,i_2}(t)]$, où $\Pi_{i_1,Q,i_2}(t)$ est la fréquence au temps t des haplotypes composés des allèles i_1 et i_2 des marqueurs flanquants et de l'allèle Q du QTL. Cette expression est dérivée dans la section "Appendix" du chapitre 8. Nous revenons ici sur certaines hypothèses faites lors de ce calcul, et essayons de montrer comment elles se justifient. Puis nous proposons quelques pistes pour étendre ce même calcul à des modèles de population plus généraux.

9.1 Hypothèses sur les fréquences des haplotypes aux marqueurs

Le calcul de l'espérance des fréquences d'haplotypes présenté dans la section "Appendix" du chapitre 8 repose sur deux hypothèses : les fréquences marginales des allèles aux marqueurs sont déterministes, et elles sont en équilibre dans le sens où

$$\Pi_{i_1,i_2}(t) = \Pi_{i_1}(t)\Pi_{i_2}(t)$$

ce que l'on peut ensuite écrire simplement $\Pi_{i_1}\Pi_{i_2}$, puisqu'il est clair (cf section 2.1.1) que dans un modèle déterministe et en l'absence de sélection et de mutation les fréquences alléliques sont constantes au cours du temps. Les hypothèses ci-dessus nous permettent donc d'écrire

$$\mathbb{E}[\Pi_{i_1,i_2}(t)\Pi_Q(t)] = \Pi_{i_1}\Pi_{i_2}\mathbb{E}[\Pi_Q(t)] = \Pi_{i_1}\Pi_{i_2}\Pi_Q(0)$$

et on obtient alors plus facilement une relation de récurrence entre $\mathbb{E}[\Pi_{i_1,Q,i_2}(t+1)]$ et $\mathbb{E}[\Pi_{i_1,Q,i_2}(t)]$.

Supposer que les fréquences des allèles aux marqueurs sont en équilibre n'est pas très grave car pour un modèle déterministe à deux loci séparés par un taux de recombinaison c on sait que

$$\Pi_{i_1,i_2}(t) = \Pi_{i_1}\Pi_{i_2} + (1-c)^t(\Pi_{i_1,i_2}(0) - \Pi_{i_1}\Pi_{i_2})$$

Ici $c = c_1 + c_2$, où c_1 est le taux de recombinaison entre le QTL et le premier marqueur et c_2 est le taux de recombinaison entre le QTL et le second marqueur (on peut supposer que ces taux de recombinaison s'ajoutent car ils sont très faibles dans le cas de la localisation fine de QTL). Cette formule montre que l'équilibre des fréquences est de manière générale très vite atteint, d'autant plus que les marqueurs sont éloignés. Ce problème a déjà été discuté dans la section 3.2, quand nous avons montré la décroissance géométrique de $\mathbb{E}[D_{i_1,i_2}(t)]$ sous un modèle de Wright-Fisher à deux loci.

L'hypothèse du déterminisme des fréquences alléliques aux marqueurs est en fait plus importante. Elle peut cependant être justifiée par le fait que la fréquence de l'allèle le plus rare d'un QTL est généralement plus faible que la fréquence de l'allèle le plus rare d'un marqueur, dans la mesure où l'allèle le plus rare du QTL est supposé être apparu plus récemment. Or pour un allèle rare, le rapport entre la variance et l'espérance de la fréquence est beaucoup plus important que pour un allèle fréquent. En effet, sous le modèle de Wright-Fisher à un locus, on a (voir section 2.1.1)

$$\mathrm{Var}(\Pi(t)) = (1-\frac{1}{2N})^t\mathrm{Var}(\Pi(0)) + \Pi(0)(1-\Pi(0))(1-(1-\frac{1}{2N})^t)$$

Dans cette expression, supposons que le premier des deux termes est nul. On obtient alors

$$\frac{\sqrt{\mathrm{Var}(\Pi(t))}}{\mathbb{E}[\Pi(t)]} = \sqrt{\frac{1-\Pi(0)}{\Pi(0)}}\sqrt{(1-(1-\frac{1}{2N})^t)}$$

Prenons par exemple une population de taille $N = 200$. Pour $\Pi(0) = 0.4$, ce rapport vaut 0.19 au bout de 10 générations, 0.42 au bout de 50 générations et 0.58 au bout de 100 générations. La variance n'est certes pas négligeable mais elle reste en tout cas relativement faible pour des temps courts. Pour $\Pi(0) = 0.1$, le rapport est de 0.47, 1.03 et 1.41 pour les mêmes temps. Les allèles rares ont donc des fréquences plus variables. Cependant, se débarrasser de l'hypothèse du déterminisme des fréquences alléliques aux marqueurs constitue un objectif intéressant pour prolonger les travaux du chapitre 8.

9.2 Extensions possibles du modèle

Dans la section "Discussion" du chapitre 8, nous avons évoqué plusieurs possibilités d'extensions de notre méthode pour prendre en compte des modèles de population un peu plus complexes. Nous revenons ici plus précisément sur certaines de ces possibilités.

9.2.1 Haplotypes flanquants

Étendre notre méthode pour tenir compte simultanément de l'information de plus de deux marqueurs est en principe assez simple. Il suffit de considérer, en chaque position, des haplotypes flanquants au lieu de marqueurs flanquants. Autrement dit, on remplace dans l'expression finale de $\mathbb{E}[\Pi_{i_1,Q,i_2}(t)]$ les termes Π_{i_1} et Π_{i_2} par Π_{h_1} et Π_{h_2}, où h_1 et h_2 sont deux haplotypes définis pour un nombre quelconque de marqueurs, h_1 étant constitué de marqueurs à gauche de x et h_2 de marqueurs à droite de x.

Le problème est ensuite de calculer Π_{h_1} et Π_{h_2}. Si on suppose comme précédemment que les marqueurs sont en équilibre, alors il n'y qu'à calculer le produit des fréquences pour les allèles composant l'haplotype. Inversement, si les marqueurs sont très proches, on peut supposer qu'il n'y a pas eu beaucoup de recombinaisons affectant ces haplotypes au cours de l'histoire. On estime donc la fréquence des haplotypes par la fréquence observée dans l'échantillon. Pour envisager des solutions intermédiaires, il faut représenter Π_{h_1} et Π_{h_2} comme des quantités pouvant dépendre du temps. Ceci complique les équations de récurrence entre $\mathbb{E}[\Pi_{h_1,Q,h_2}(t+1)]$ et $\mathbb{E}[\Pi_{h_1,Q,h_2}(t)]$, mais la résolution est peut être encore possible.

Il serait intéressant de tester si ces différentes options permettent d'améliorer la précision des estimations par rapport à la méthode actuelle.

9.2.2 Prise en compte de la mutation

Le modèle de Wright-Fisher à trois loci utilisé au chapitre 8 peut être modifié pour tenir compte de la possibilité de mutations entre allèles au niveau de chaque locus. Trouver une expression pour $\mathbb{E}[\Pi_{i_1,Q,i_2}(t)]$ devient alors plus difficile, mais le calcul reste sans doute faisable moyennant certaines hypothèses sur la structure des mutations. Nous en donnons un aperçu en démarrant le calcul de $\mathbb{E}[\Pi_{i,Q}(t)]$ pour un modèle à deux loci avec mutation, étendant ainsi les calculs de la section "Appendix" du chapitre 8.

Reprenons donc les hypothèses et notations de cette section. Introduisons de plus, à chaque génération, une probabilité μ_Q de mutation de l'allèle q vers l'allèle Q, et une probabilité μ_q de mutation de l'allèle Q vers l'allèle q. Pour le marqueur, on introduit de

même la matrice de mutation $\mu = (\mu_{i_1,i_2})_{1\leq i_1,i_2\leq I}$, où μ_{i_1,i_2} $(i_1 \neq i_2)$ est la probabilité de mutation de l'allèle i_1 vers l'allèle i_2, et où $\mu_{i_1,i_1} = 1 - \sum_{i_2\neq i_1} \mu_{i_1,i_2}$.

La probabilité de tirer un haplotype (i_1, Q) à la génération $(t+1)$ s'écrit désormais

$$r_{i_1,Q}(t) = (1-c)\Pi^*_{i_1,Q}(t) + c\Pi^*_Q(t)\Pi^*_{i_1}(t)$$

où les fréquences $\Pi^*_{i_1,Q}(t)$, $\Pi^*_Q(t)$ et $\Pi^*_{i_1}(t)$ sont des fréquences virtuelles dans la population après que les mutations soient intervenues. On a

$$\begin{aligned}
\Pi^*_{i_1,Q}(t) &= \sum_{i_2=1}^{I} \mu_{i_2,i_1}(\mu_Q \Pi_{i_2,q}(t) + (1-\mu_q)\Pi_{i_2,Q}(t)) \\
\Pi^*_Q(t) &= \mu_Q \Pi_q(t) + (1-\mu_q)\Pi_Q(t) \\
\Pi^*_{i_1}(t) &= \sum_{i_2=1}^{I} \mu_{i_2,i_1}\Pi_{i_2}(t)
\end{aligned}$$

Dans le cadre du modèle de Wright-Fisher, on sait que $\mathbb{E}[\Pi_{i_1,Q}(t+1)] = \mathbb{E}[r_{i_1,Q}(t)]$. En supposant, comme dans le cas sans mutation, que les fréquences des marqueurs sont déterministes, cela donne

$$\begin{aligned}
\mathbb{E}[\Pi_{i_1,Q}(t+1)] &= (1-c)\sum_{i_2=1}^{I} \mu_{i_2,i_1}(\mu_Q \mathbb{E}[\Pi_{i_2,q}(t)] + (1-\mu_q)\mathbb{E}[\Pi_{i_2,Q}(t)]) \\
&+ c(\mu_Q \mathbb{E}[\Pi_q(t)] + (1-\mu_q)\mathbb{E}[\Pi_Q(t)])(\sum_{i_2=1}^{I} \mu_{i_2,i_1}\Pi_{i_2}(t))
\end{aligned}$$

On peut arranger cette formule en remarquant que $\Pi_q(t) = 1 - \Pi_Q(t)$ et que $\Pi_{i_2,q}(t) = \Pi_{i_2}(t) - \Pi_{i_2,Q}(t)$. On obtient finalement

$$\begin{aligned}
\mathbb{E}[\Pi_{i_1,Q}(t+1)] &= \mu_Q \sum_{i_2=1}^{I} \mu_{i_2,i_1}\Pi_{i_2}(t) \\
&+ (1-c)(1-\mu_Q-\mu_q)(\sum_{i_2=1}^{I} \mu_{i_2,i_1}\mathbb{E}[\Pi_{i_2,Q}(t)]) \\
&+ c(1-\mu_Q-\mu_q)(\sum_{i_2=1}^{I} \mu_{i_2,i_1}\Pi_{i_2}(t))\ \mathbb{E}[\Pi_Q(t)]
\end{aligned}$$

Notons $\Pi(t) = (\Pi_{1,Q}(t), \ldots, \Pi_{I,Q}(t))$. On obtient alors l'écriture matricielle

$$\mathbb{E}[\Pi(t+1)] = \mu_Q(\Pi_1(t), \ldots, \Pi_I(t))\mu + (1-\mu_Q-\mu_q)\mathbb{E}[\Pi(t)](cA + (1-c)Id_I)\mu$$

où $A = (\Pi_1, \cdots, \Pi_I) \otimes \mathbb{1}_I$, $\otimes$ étant le produit de Kronecker pour les matrices, $\mathbb{1}_I$ est le vecteur colonne de taille I dont toutes les coordonnées valent 1, et Id_I est la matrice identité de taille $I \times I$. On retrouve bien l'expression de la section "Appendix" du chapitre 8 si $\mu_Q = \mu_q = 0$, $\mu = Id_I$ et si les fréquences $\Pi_i(t)$ sont invariantes au cours du temps. Notons qu'ici cette hypothèse n'est pas aussi naturelle que dans le modèle sans mutation. Cependant, pour faciliter les calculs, on peut supposer qu'elle est vérifiée dans la mesure où les fréquences alléliques tendent vers un équilibre stable, comme l'a illustré la section 2.1.1.

Que l'on suppose ou pas que les fréquences des marqueurs sont constantes au cours du temps, trouver une expression de $\mathbb{E}[\Pi(t)]$ en fonction de $\mathbb{E}[\Pi(0)]$ et des autres paramétres du modèle implique de pouvoir exprimer simplement le terme $((cA + (1-c)Id_I)\mu)^t$ pour tout $t \in \mathbb{N}$. On a vu au chapitre 8 que la matrice A était idempotante, ce qui facilite un peu les choses. On remarque également que $\mu A = A$. Pour que le calcul soit faisable, il faudrait cependant avoir une expression simple de μ^t. Un choix judicieux de la matrice μ doit permettre d'obtenir un résultat. Mais il reste encore du travail ensuite pour exprimer l'espérance des fréquences d'haplotypes dans le modèle à trois loci.

9.2.3 Prise en compte de la sélection

Une autre extension intéressante serait de pouvoir traiter les cas où le QTL est soumis à des phénomènes de sélection. Comme dans la section précédente, reprenons le cadre du modèle à deux loci utilisé dans la section "Appendix" du chapitre 8 et voyons ce que change la prise en compte de la sélection. On a à nouveau

$$r_{i,Q}(t) = (1-c)\Pi^*_{i,Q}(t) + c\Pi^*_Q(t)\Pi_i(t)$$

Mais cette fois les fréquences $\Pi^*_{i,Q}(t)$ et $\Pi^*_Q(t)$ sont les fréquences virtuelles après sélection. Les allèles du marqueur ne sont pas sujets à la sélection. Conformément à ce que nous avions fait en 2.1.1, nous introduisons les viabilités $\omega_{Q/Q}$, $\omega_{Q/q}$ et $\omega_{q/q}$ des génotypes Q/Q, Q/q et q/q. Elles nous permettent d'exprimer

$$\begin{aligned}
\Pi^*_Q(t) &= \frac{\omega_{Q/Q}\Pi^2_Q(t) + \omega_{Q/q}\Pi_Q(t)\Pi_q(t)}{\omega_{Q/Q}\Pi^2_Q(t) + 2\omega_{Q/q}\Pi_Q(t)\Pi_q(t) + \omega_{q/q}\Pi^2_q(t)} \\
\Pi^*_{i,Q}(t) &= \frac{\omega_{Q/Q}\Pi_{i,Q}(t)\Pi_Q(t) + \omega_{Q/q}\Pi_{i,Q}(t)\Pi_q(t)}{\omega_{Q/Q}\Pi^2_Q(t) + 2\omega_{Q/q}\Pi_Q(t)\Pi_q(t) + \omega_{q/q}\Pi^2_q(t)}
\end{aligned}$$

Le fait qu'il y ait des quotients dont le dénominateur implique $\Pi_Q(t)$ rend difficile le raisonnement par récurrence que nous utilisions pour des modèles neutres. Il ne sera donc pas très aisé de tenir compte de la sélection dans notre méthode.

Quatrième partie

Recherche d'associations dans une population structurée

Chapitre 10

Recherche d'associations dans une population structurée

Dans les parties II et III de ce manuscrit, nous nous sommes intéressés au problème de l'estimation fine de la position d'un QTL. Une autre approche, de plus en plus utilisée en cartographie, consiste à tester directement, pour un très grand nombre de loci répartis sur le génome, si ces loci sont ou pas des QTL pour le caractère étudié. Pour répondre à cette question, on utilise des méthodes d'association telles que celles présentées en 4.2. Dans le cadre d'un projet de détection de QTL nommé GeMqual, j'ai eu l'occasion d'étudier les propriétés d'une de ces méthodes, le *Transmission Disequilibrium Test*, qui est utilisé dans le cas où la population est structurée. Les résultats obtenus sont exposés dans ce chapitre.

Nous précisons dans un premier temps le contexte de ce travail, à savoir les méthodes d'association, les problèmes rencontrés par ces méthodes dans le cas d'une population structurée et enfin les grandes lignes du projet GeMqual. Nous en venons ensuite au travail réalisé.

10.1 Contexte du travail

10.1.1 Méthodes d'association

Les méthodes d'association ont déjà été introduites en 4.2. L'objectif était alors d'estimer la position la plus probable d'un QTL. Dans le cadre de ce chapitre, nous nous en tenons à leur usage initial, qui est de détecter les zones du génome ou il y a des QTL. Nous rappelons ici le principe général de ces méthodes.

On observe, pour un échantillon de N_s individus, les phénotypes z_n pour un carac-

tère quantitatif donné, et les génotypes m_n au niveau d'un marqueur. Supposons que le marqueur a I allèles; il y a donc $K = \frac{I(I-1)}{2}$ génotypes possibles pour ce marqueur. On considère le modèle linéaire :

$$Z_{k,j} = m_k + \epsilon_{k,j},\ k = 1, \ldots, K,\ j = 1, \ldots, N_k\ (M)$$

où les $\epsilon_{k,j}$ sont des bruits gaussiens indépendants et où N_k est le nombre d'individus ayant le génotype k. On cherche à tester si le marqueur est lié à un QTL, c'est à dire s'il existe un QTL dont le taux de recombinaison c avec le marqueur est strictement inférieur à $\frac{1}{2}$. On teste pour cela l'hypothèse H_0 : "$m_1 = \ldots = m_K$". On dit que le marqueur est associé au QTL si H_0 est rejetée.

Si le marqueur est exactement le QTL, les phénotypes observés dépendent directement des génotypes au marqueur, suivant le modèle décrit par l'équation (1.1). H_0 est donc très probablement rejetée. De manière générale, H_0 est aussi rejetée s'il y a du déséquilibre de liaison entre le marqueur et un QTL. En effet, supposons par exemple que l'allèle Q du QTL donne des plus grandes valeurs de phénotypes. Dans ce cas, les allèles i du marqueur qui sont positivement associés à l'allèle Q (c'est à dire pour lesquels $D_{i,Q} > 0$, où $D_{i,Q}$ est la mesure de déséquilibre définie par l'équation (3.1)), correspondent plus souvent que les autres à des grandes valeurs de phénotypes. Ceci engendre donc en principe des écarts entre les moyennes des phénotypes observés dans chaque groupe de génotypes au marqueur. D'autre part, nous avons montré en 3.2 que dans une population homogène, un déséquilibre de liaison important entre deux loci ne peut être expliqué que par un taux de recombinaison faible entre ces loci. Par conséquent le test de H_0 permet bien de tester si le marqueur est lié ou pas à un QTL.

Ce test est généralement effectué successivement pour un très grand nombre de marqueurs répartis sur le génome, de façon à repérer les zones du génome ayant une influence sur un caractère donné. Ce type de démarches porte le nom d'*étude d'associations*. Un problème majeur de ces études est de contrôler le taux de faux positifs, puisque de l'argent et du temps de recherche sont généralement investis pour étudier plus précisément les zones détectées par ces tests. Or, deux phénomènes majeurs sont à prendre en compte à cet égard :

- la multiplicité des tests : le nombre de marqueurs analysés est souvent de l'ordre de plusieurs centaines, et peut aller jusqu'à plusieurs centaines de milliers pour les études les plus importantes. Pour ajuster le seuil des tests, une solution classique consiste à utiliser la correction de Bonferroni. Cependant la perte de puissance est alors importante, car on sait que pour des marqueurs proches les tests sont corrélés.

Parmi les alternatives couramment utilisées, on peut citer les méthodes de contrôle du taux de faux positifs (Benjamini et Hochberg, 1995; Weller et al., 1998) ou les méthodes de permutation (McIntyre et al., 2000). Plus récemment, différentes stratégies consistant à identifier des groupes de marqueurs corrélés et à les tester simultanément se sont développées. Dans chaque groupe de marqueurs, on peut par exemple évaluer les corrélations entre tests en mesurant le déséquilibre de liaison entre marqueurs (Nyholt, 2004), ou effectuer directement un test sur les haplotypes, comme cela a été expliqué en 4.2.

– la structure des populations : si la population considérée n'est pas homogène, il est fréquent de détecter qu'un marqueur est associé alors qu'il n'est en fait lié à aucun QTL. Les causes de ce phénomène, et les solutions adoptées, sont décrites dans la section suivante.

10.1.2 Problème de la structure

Revenons sur le modèle (M) décrit dans la section précédente. Nous avons vu que dans ce modèle, on observait des différences entre les moyennes phénotypiques μ_j à condition que certains allèles du marqueur soient en déséquilibre de liaison avec les allèles d'un QTL. Nous avons aussi rappelé que dans une population homogène, ce déséquilibre de liaison ne peut être expliqué que par un taux de recombinaison faible avec le QTL. Mais dans une population structurée, on peut observer du déséquilibre de liaison entre les allèles de loci très éloignés, simplement du fait des différences de fréquences alléliques entre sous-populations (voir l'article de Cierco-Ayrolles et al. (2004) en annexe). On parle alors, un peu abusivement, de faux positifs pour désigner les marqueurs associés au caractère (pour lesquels H_0 est rejetée) alors qu'ils ne sont liés à aucun QTL. Pour contrôler le nombre de ces faux positifs, trois types d'approches ont été proposés.

La première approche porte le nom de *contrôle génomique*, et a été proposée par Devlin et Roeder (1999). Elle est a priori plutôt adaptée à des caractères discrets. Par exemple, si on recherche les marqueurs liés à une maladie, on compare généralement les effectifs des génotypes au marqueur entre les individus malades et non malades, et on teste la différence entre ces effectifs à l'aide d'un test du chi-deux. Dans une population hétérogène, la distribution de la statistique de test sous l'hypothèse nulle (le marqueur n'est lié à aucun QTL) n'est pas le chi-deux attendu. Mais Devlin et Roeder (1999) ont montré que sous certaines hypothèses, cette distribution ne différait de celle attendue que par un facteur multiplicatif λ. Ils ont donc proposé d'estimer préalablement ce paramètre en calculant la statistique pour un ensemble de marqueurs indépendants dont on sait

qu'ils ne sont pas liés à la maladie.

Une autre solution peut être d'estimer la structure inconnue de la population. La méthode proposée par Pritchard, Stephens, et Donnelly (2000) est une des plus connues dans ce domaine. Elle suppose que les individus observés sont issus de r populations distinctes, dans lesquelles il n'y a pas de déséquilibre de liaison entre loci. Chaque population est caractérisée par la donnée des fréquences alléliques pour L loci. L'observation des génotypes au niveau de ces loci pour un échantillon d'individus permet d'estimer toutes les fréquences alléliques, et les probabilités pour chaque individu d'appartenir à une population donnée. Cette méthode s'appuie sur les similarités de génotypes entre individus. Elle utilise un cadre Bayésien et un algorithme MCMC. Pour détecter les marqueurs liés à un caractère dans le cas d'une population structurée, on peut donc commencer par estimer ainsi l'origine de chaque individu en utilisant un ensemble de marqueurs indépendants, puis tester l'association des marqueurs population par population. Pritchard, Stephens, Rosenberg, et Donnelly (2000) ont mis en application cette stratégie pour détecter des gènes de maladie, mais on peut facilement l'adapter au cas des QTL. D'autres méthodes similaires sont présentées par exemple par Köhler et Bickeböller (2005).

Une dernière stratégie consiste à génotyper également les parents des individus dont on observe le phénotype, et à étudier la transmission des allèles des parents aux enfants. On étudie donc des familles de type père-mère-enfant communément appelées *trios*. Cette stratégie a été initialement utilisée par Spielman et al. (1993) pour localiser des gènes de maladie à l'aide de marqueurs bialléliques. Elle porte le nom de TDT (pour "Transmission-Desiquilibrium Test"). De nombreuses extensions ont été proposées depuis, permettant d'incorporer notamment des structures de familles ou des génotypes plus complexes. Allison (1997) a également étendu cette méthode à l'étude de caractères quantitatifs, en décrivant plusieurs tests basés sur le principe suivant. Soit un marqueur biallélique, et un couple parent-enfant pour lequel le parent est hétérozygote 1/2. On note Z le phénotype de l'enfant et T l'allèle qui lui est transmis par ce parent. Sous l'hypothèse nulle d'absence de liaison du marqueur avec un QTL, on a

$$\mathbb{E}[Z \mid T = 1] = \mathbb{E}[Z \mid T = 2]$$

et ce indépendamment de la structure de la population. On peut donc rejeter l'hypothèse nulle si on constate parmi les observations un écart par rapport à ces résultats. Notons toutefois que la seule liaison ne suffit pas à rejeter l'hypothèse nulle : il faut aussi qu'il y ait du déséquilibre de liaison entre le marqueur et le QTL (d'où le nom de TDT). En effet, notons Q et q les allèles du QTL. et G celui de ces allèles qui est transmis par le

parent dans le couple parent-enfant présenté ci-dessus. On a

$$\begin{aligned}\mathbb{E}[Z \mid T=1] &= \mathbb{E}[Z \mid G=Q]\ \mathbb{P}(G=Q \mid T=1) + \mathbb{E}[Z \mid G=q]\ \mathbb{P}(G=q \mid T=1)\\ \mathbb{E}[Z \mid T=2] &= \mathbb{E}[Z \mid G=Q]\ \mathbb{P}(G=Q \mid T=2) + \mathbb{E}[Z \mid G=q]\ \mathbb{P}(G=q \mid T=2)\end{aligned}$$

Si, dans la sous-population dont est issu le couple parent-enfant étudié, il n'y a pas de déséquilibre de liaison entre le marqueur et le QTL, on a

$$\mathbb{P}(G=Q \mid T=1) = \mathbb{P}(G=Q \mid T=2)$$

et

$$\mathbb{P}(G=q \mid T=1) = \mathbb{P}(G=q \mid T=2)$$

Par conséquent on voit que

$$\mathbb{E}[Z \mid T=1] = \mathbb{E}[Z \mid T=2]$$

même si $\mathbb{E}[Z \mid G=Q]$ et $\mathbb{E}[Z \mid G=q]$ sont très différents. Inversement, plus le déséquilibre de liaison est important, plus la puissance du test sera importante.

Abecassis et al. (2000) ont proposé un modèle linéaire permettant de prendre en compte l'idée de transmission contenue dans le TDT. Nous étudierons en 10.2 la puissance d'un test d'association dérivant de ce modèle.

10.1.3 Projet GeMqual

Le travail présenté dans ce chapitre a été réalisé dans le cadre d'un projet européen nommé GeMqual (http ://www.gemqual.org), dont le but est de détecter des gènes ayant une influence sur la qualité de la viande bovine. Pour cette étude, 450 trios issus de 15 races de bovins viande d'Europe occidentale (exactement 30 par race) ont été observés. Pour chaque veau, une centaine de phénotypes liés à la qualité de la viande ont été mesurés. Certains phénotypes étaient de nature chimique (taux d'un certain acide dans la viande ...) ou physique (poids, taille de l'animal, ...). D'autres provenaient de tests gustatifs réalisés sur des panels d'individus volontaires. Tous les animaux (veaux, vaches et taureaux) ont été génotypés pour quelque 300 marqueurs biallélіques (des SNP) répartis dans une centaine de gènes candidats (c'est à dire sélectionnés par les biologistes car susceptibles d'être impliqués dans la qualité de la viande) répartis sur tout le génome bovin.

Une des nombreuses questions soulevées par ce projet concerne naturellement le choix

de la méthode statistique à utiliser pour analyser les données et détecter les loci ayant une influence sur les phénotypes. Quand ce projet a commencé à se mettre en place il y a quelques années, il était prévu d'utiliser une méthode de type TDT, et c'est pour cela que la structure en trios a été choisie. En fait, beaucoup de parents n'ont pas pu être identifiés avec certitude, si bien qu'à l'issue des analyses le génotype des deux parents à la fois n'était disponible que pour la moitié des veaux. D'autre part, les données de GeMqual ont ceci de particulier que la structure en races, qui est probablement la structure principale dans les données de génotype, est connue. Nous avons donc pensé qu'un modèle tenant compte explicitement de la structure en races, et n'utilisant pas l'information provenant des parents, serait plus approprié. Pour valider ce choix, nous avons donc cherché à comparer de manière théorique la puissance d'un test de type TDT basé sur les 450 trios, avec celle d'une méthode testant simultanément l'association dans chacune des races. Cette comparaison est l'objet de la suite du chapitre.

10.2 Puissance asymptotique du TDT dans le cas d'un échantillon de structure connue

Soit un échantillon de trios issus de r populations distinctes. On suppose que le nombre de trios provenant de chaque population est le même. Ce nombre est noté n, et la taille totale de l'échantillon est donc rn. Pour le trio j de la population i, on mesure :

- le phénotype $z_{i,j}$ de l'enfant pour un caractère quantitatif.
- le génotype $m_{i,j}$ de l'enfant pour un marqueur biallélique d'allèles 1 et 2.
- les génotypes $p_{i,j}^{(1)}$ et $p_{i,j}^{(2)}$ des deux parents pour ce même marqueur.

À partir de ces observations, on souhaite tester si le marqueur est lié à un QTL pour le caractère. Le problème est que la structure de la population n'est a priori pas connue, ce qui implique qu'on ne sait pas de quelle population provient chaque individu. Notre objectif est en fait d'évaluer les conséquences de cette absence d'information pour la détection de la liaison.

Nous commençons pour cela par présenter le vrai modèle que sont supposées suivre les variables $Z_{i,j}$, $M_{i,j}$, $P_{i,j}^{(1)}$ et $P_{i,j}^{(2)}$. Puis nous décrivons trois tests utilisés pour détecter la liaison. Le premier, qui nous sert de référence, est construit à partir du vrai modèle et suppose que la structure de la population est connue. A l'inverse, les deux autres tests sont des tests utilisés quand la structure de la population est inconnue. L'un de ces tests est de type TDT, c'est à dire qu'il utilise les génotypes des parents. Nous insisterons plus particulièrement sur celui-ci.

Dans un cadre asymptotique local, et sous le vrai modèle, nous donnons ensuite la limite en distribution des trois tests quand n tend vers l'infini, ce qui nous permet d'étudier les erreurs de première espèce et les puissances de ces tests. Les démonstrations de convergence sont présentées en fin de chapitre. Bien que le modèle utilisé soit très proche d'un modèle linéaire classique, il diffère de celui-ci dans la mesure où les régresseurs $M_{i,j}$ sont considérés comme des variables aléatoires. Ces variables sont de plus corrélées aux $P_{i,j}$, qui interviennent dans l'expression de la statistique du TDT. C'est ce qui fait la difficulté de la démonstration.

10.2.1 Vrai modèle

Soit un trio père-mère-enfant, d'indice j, pris au hasard dans la population i. Si π_i est la fréquence de l'allèle 1 du marqueur dans cette population, on suppose que le génotype d'un parent prend les valeurs :

$$\begin{cases} 1/1 & \text{avec probabilité} \quad \pi_i^2 \\ 1/2 & \text{avec probabilité} \quad 2\pi_i(1-\pi_i) \\ 2/2 & \text{avec probabilité} \quad (1-\pi_i)^2 \end{cases}$$

Pour former le génotype de l'enfant, on choisit le génotype de ses deux parents selon la loi décrite ci-dessus, et on tire un allèle de chacun d'eux avec les probabilités $\frac{1}{2}/\frac{1}{2}$ (on suppose donc que la transmission des gènes est mendélienne). On montre facilement (Abecassis et al., 2000) que la loi marginale du génotype de l'enfant est la même que celle décrite ci-dessus pour les parents..

Si le marqueur est lié au caractère, le phénotype de cet enfant peut être représenté par le modèle (VM) (comme vrai modèle)

$$Z_{i,j} = \mu_i + a_i(M_{i,j} - \overline{M}_i) + \epsilon_{i,j},\ i = 1, \ldots, r,\ j = 1, \ldots, n$$

où μ_i est la valeur moyenne des phénotypes pour la population i, a_i est l'effet additif du marqueur dans la population i et où $\epsilon_{i,j}$ est une variable de loi $\mathcal{N}(0, \sigma^2)$. Les $\epsilon_{i,j}$ sont indépendantes entre elles ainsi que des $M_{i,j}$. Dans cette formule, les génotypes sont codés de la manière suivante :

$$M_{i,j} = \begin{cases} 1 & \text{si le génotype est } 1/1 \\ 0 & \text{si le génotype est } 1/2 \\ -1 & \text{si le génotype est } 2/2 \end{cases}$$

et $\overline{M}_i = \frac{1}{n}\sum_{j=1}^{n} M_{i,j}$. Le modèle (VM) est donc un modèle additif pour l'effet du marqueur. Nous avons fait ce choix afin de simplifier un peu les calculs qui vont suivre mais on pourrait tout à fait ajouter un terme de dominance. En revanche, il est important de supposer a priori un effet génétique a_i distinct pour chaque population, et ce pour deux raisons. Premièrement, l'effet du QTL sur le caractère peut dépendre de l'environnement des individus, et donc de la population. Deuxièmement, nous mesurons ici l'effet du marqueur, qui peut être vu comme une fonction de l'effet du QTL et du déséquilibre de liaison entre le QTL et le marqueur (Nielsen et Weir, 1999). Il n'y aucune raison pour que ce déséquilibre soit le même dans les différentes populations. Dans le modèle (VM), l'absence de liaison entre le QTL et le marqueur correspond à l'hypothèse (H_0)

$$a_1 = \ldots = a_r = 0$$

Sous l'hypothèse (H_0), les phénotypes sont simplement décrits par le modèle

$$Z_{i,j} = \mu_i + \epsilon_{i,j},\ i = 1, \ldots, r,\ j = 1, \ldots, n$$

Le modèle (VM) suppose que le génotype des parents n'a pas d'influence directe sur le phénotype de l'enfant. Cependant, un des tests que nous allons étudier utilise les génotypes des parents, c'est pourquoi nous avons aussi décrit leur loi. La variable utilisée par le test est en fait $P_{i,j} = \frac{P_{i,j}^{(1)} + P_{i,j}^{(2)}}{2}$, où $P_{i,j}^{(1)}$ et $P_{i,j}^{(2)}$ sont les génotypes des deux parents, codés de la même manière que $M_{i,j}$. Pour le calcul de la loi asymptotique des tests, on utilisera notamment les propriétés suivantes :

$$\mathbb{E}[M_{i,j}] = \mathbb{E}[P_{i,j}] = 2\pi_i - 1$$

$$\mathrm{Var}(M_{i,j}) = 2\pi_i(1 - \pi_i)$$

$$\mathrm{Var}(P_{i,j}) = \mathrm{Cov}(M_{i,j}, P_{i,j}) = \pi_i(1 - \pi_i)$$

10.2.2 Modèles d'analyse et statistiques utilisées

Structure de population connue

Si on connaît la structure de la population, on peut construire un test de liaison à partir du modèle (VM). Il s'agit du test de l'hypothèse H_0 : "$a_1 = \ldots = a_r = 0$", basé sur la statistique de Fisher

$$\hat{F} = \frac{(SCR_0 - SCR_1)/r}{SCR_1/(rn - r)}$$

où SCR_1 est la somme des carrés résiduels sous l'hypothèse (H_1) et SCR_0 est la somme des carrés résiduels sous l'hypothèse (H_0). Nous utiliserons par la suite les notations vectorielles

$$(H_1) \quad Z = X\ \theta + \epsilon$$

et

$$(H_0) \quad Z = \underline{X}\ \underline{\theta} + \epsilon$$

où $Z = {}^t(Z_{1,1}, Z_{1,2}, \ldots, Z_{r,n})$, $\epsilon = {}^t(\epsilon_{1,1}, \epsilon_{1,2}, \ldots, \epsilon_{r,n})$, $\theta = {}^t(\mu_1, \ldots, \mu_r, a_1, \ldots, a_r)$ et $\underline{\theta} = {}^t(\mu_1, \ldots, \mu_r)$. X et $\underline{X}$ sont des matrices d'incidence de tailles respectives $rn \times 2r$ et $rn \times r$. On suppose qu'elles sont réversibles, ce qui implique que dans chaque race i $(i \in 1, \ldots, r)$ il y ait au moins deux génotypes différents.

Introduisons enfin P_X et $P_{\underline{X}}$, les projecteurs orthogonaux sur les espaces vectoriels engendrés par les colonnes de X et les colonnes de $\underline{X}$. Ces espaces vectoriels sont notés E_X et $E_{\underline{X}}$. Avec ces notations on a :

$$SCR_1 = ||Z - P_X Z||^2$$

et

$$SCR_0 = ||Z - P_{\underline{X}} Z||^2$$

Ce test est un test standard, et on sait que sous (H_0) et quand n tend vers l'infini $r\hat{F}$ suit une loi du chi-deux à r degrés de liberté. On rejette donc (H_0) pour

$$r\hat{F} \notin [q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]$$

où $q_{\frac{\alpha}{2}}$ et $q_{1-\frac{\alpha}{2}}$ sont les quantiles $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$ de la loi χ^2_r.

Structure de population inconnue

Supposons maintenant que la structure de la population est inconnue. Autrement dit on ne sait pas de quelle sous-population provient chaque individu. Dans ce cas, une première idée est d'utiliser le modèle d'analyse (MA_1)

$$Z_{i,j} = \mu_0 + a_0(M_{i,j} - \overline{M}) + \epsilon_{i,j},\ i = 1, \ldots, r,\ j = 1, \ldots, n$$

où μ_0 est la valeur moyenne des phénotypes, a_0 est l'effet additif du marqueur et $\overline{M} = \frac{1}{nr} \sum_{i=1}^{r} \sum_{j=1}^{n} M_{i,j}$. Ce modèle n'est utilisé que pour construire le test de liaison entre le marqueur et le QTL. Il s'agit ici du test de l'hypothèse H_0 : "$a_0 = 0$" basé sur la

statistique de Student

$$\hat{T}_1 = \frac{\hat{a}_0}{\sqrt{\hat{\sigma}^2}} \sqrt{\sum_{i=1}^{r} \sum_{j=1}^{n} (M_{i,j} - \overline{M})^2}$$

où

$$\hat{a}_0 = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n} Z_{i,j}(M_{i,j} - \overline{M})}{\sum_{i=1}^{r} \sum_{j=1}^{n} (M_{i,j} - \overline{M})^2}$$

est l'estimateur des moindres carrés de a_0 et où $\hat{\sigma}^2$ est l'estimateur des moindres carrés de σ^2. Si (MA_1) était le vrai modèle, $\hat{T}_1$ serait, sous H_0 : "$a_0 = 0$" et quand n tend vers l'infini, de loi normale $\mathcal{N}(0,1)$. On rejette donc ici (H_0) si

$$\hat{T}_1 \notin [-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$.

Si on pense que la population est structurée, mais qu'on ne connaît pas pour autant cette structure, une deuxième solution consiste à utiliser les génotypes des parents pour obtenir un test plus fiable (nous préciserons dans quel sens dans la section suivante). Par exemple, Abecassis et al. (2000) ont proposé le modèle d'analyse $(MA_2$:

$$Z_{i,j} = \mu_0 + \beta_b(P_{i,j} - \overline{P}) + \beta_w(M_{i,j} - P_{i,j}) + \epsilon_{i,j},\ i = 1, \ldots, r,\ j = 1, \ldots, n$$

où $\overline{P} = \frac{1}{rn} \sum_{i=1}^{r} \sum_{j=1}^{n} P_{i,j}$. β_b représente l'effet entre familles ("b" comme "between") et β_w l'effet intra-familles ("w" comme "within"). En notation vectorielle, ce modèle s'écrit

$$(MA_2) \quad Z = \widetilde{X}\widetilde{\theta} + \epsilon$$

où $\widetilde{\theta} = {}^t(\mu_0, \beta_b, \beta_w)$ et où $\widetilde{X}$ est une matrice d'incidence supposée inversible. On pose $\widetilde{X} = (X_{\mu_0}, X_{\beta_b}, X_{\beta_w})$, où X_{μ_0}, X_{β_b} et X_{β_w} sont des vecteurs de taille $rn \times 1$. Ce modèle n'est utilisé que pour construire le test de liaison entre le marqueur et le QTL. Il s'agit ici du test de l'hypothèse H_0 : "$\beta_w = 0$" basé sur la statistique de Student

$$\hat{T}_2 = \frac{\hat{\beta}_w}{\sqrt{\hat{\sigma}^2} \sqrt{e_3({}^t\widetilde{X}\widetilde{X})^{-1}({}^t e_3)}}$$

où $e_3 = (0, 0, 1)$ et où $\hat{\beta}_w$ et $\hat{\sigma}^2$ sont les estimateurs des moindres carrés de β_w et σ^2 dans

le modèle $(\widetilde{M})$:

$$\begin{aligned}\hat{\beta}_w &= e_3({}^t\widetilde{X}\widetilde{X})^{-1}({}^t\widetilde{X}Z) \\ \hat{\sigma}^2 &= \frac{||Z - P_{\widetilde{X}}Z||^2}{rn-3}\end{aligned}$$

Si (MA_2) était le vrai modèle, $\hat{T}_2$ serait, sous H_0 : "$\beta_w = 0$" et quand n tend vers l'infini, de loi normale $\mathcal{N}(0,1)$. On rejette donc ici (H_0) si

$$\hat{T}_2 \notin [-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]$$

Dans la suite, nous appellerons souvent TDT, pour "Transmission Disequilibrium Test", le test $\hat{T}_2$. Tout comme le vrai TDT décrit à l'origine par Spielman et al. (1993), il s'intéresse en effet, à travers le prédicteur $P_{i,j} - M_{i,j}$, à la transmission des gènes des parents à l'enfant.

10.2.3 Loi asymptotique des statistiques de test

Pour étudier les qualités des trois tests que nous venons d'introduire, nous présentons ici un résultat de convergence en distribution pour le trois statistiques quand r est fixe et n tend vers l'infini. Pour obtenir cette convergence, nous considérons le modèle local (VM_n)

$$Z_{i,j} = \mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i) + \epsilon_{i,j},\ i = 1, \ldots, r,\ j = 1, \ldots, n$$

qui est identique à (VM) mais dans lequel les effets du marqueur dans chaque population sont d'ordre $\frac{1}{\sqrt{n}}$. Il s'agit du modèle local classique utilisé dans le modèle linéaire pour obtenir des puissances de tests qui ne tendent pas vers l'infini.

Proposition 2 *Si les variables $M_{i,j}$ et $P_{i,j}$ suivent la loi décrite en 10.2.1 et si la loi des variables $Z_{i,j}$ est déterminée par le modèle (VM_n), alors :*

1. $$\hat{F} \xrightarrow{\mathcal{L}} \frac{1}{r}\chi^2_r\left(\frac{2}{\sigma^2}\sum_{i=1}^{r} u_i^2\pi_i(1-\pi_i)\right)$$

   où χ^2_r est un chi-deux décentré à r degrés de liberté.

2. *A moins que $\mu_1 = \ldots = \mu_r$ ou $\pi_1 = \ldots = \pi_r$,*

   $$\hat{T}_1 \xrightarrow{p} +\infty$$

3.

$$\hat{T}_2 \xrightarrow{\mathcal{L}} \mathcal{N}\left(\frac{\sum_{i=1}^{r} u_i \pi_i (1-\pi_i)}{\sqrt{(\sigma^2+l)\sum_{i=1}^{r} \pi_i (1-\pi_i)}}, \frac{\sum_{i=1}^{r} \pi_i (1-\pi_i)(\mu_i^2+\sigma^2)}{(\sigma^2+l)\sum_{i=1}^{r} \pi_i (1-\pi_i)}\right)$$

où

$$\begin{aligned} l1 &= \frac{1}{r}\sum_{i=1}^{r} \mu_i^2 \\ l2 &= \frac{1}{r^2}\left(\sum_{i=1}^{r} \mu_i\right)^2 \\ l3 &= \frac{\left(\frac{1}{r}\sum_{i=1}^{r}(\mu_i(2\pi_i-1)) - \frac{1}{r^2}\left(\sum_{i=1}^{r}\mu_i\right)\left(\sum_{i=1}^{r}(2\pi_i-1)\right)\right)^2}{\frac{1}{r}\sum_{i=1}^{r}(2\pi_i-1)^2 + \frac{1}{r}\sum_{i=1}^{r}\pi_i(1-\pi_i) - \frac{1}{r^2}\left(\sum_{i=1}^{r}(2\pi_i-1)\right)^2} \end{aligned}$$

La preuve de cette proposition est donnée en section 10.2.6.

10.2.4 Interprétation

Sous le modèle (VM_n) dans lequel la population est structurée, la statistique $\hat{T}_1$ tend vers l'infini aussi bien sous (H_0) que sous (H_1). Ce test n'est donc pas valide puisqu'il conduit asymptotiquement à toujours rejeter l'hypothèse nulle. Ce résultat n'est pas nouveau, c'est même ce qui a motivé l'utilisation des tests de type TDT. Dans la preuve de la section 10.2.6, il apparaît clairement que ce problème de convergence vient de l'hétérogénéité des fréquences et des valeurs moyennes des phénotypes entre les population. Intuitivement, cela peut se comprendre simplement de la manière suivante : si dans une sous-population les valeurs de phénotypes sont importantes (sous l'effet de l'environnement ou de gènes indépendants du marqueur) et si parrallélement la fréquence d'un allèle du marqueur dans cette sous-population est supérieure à la fréquence moyenne dans la population globale, une analyse qui ne tient pas compte de la structure en sous-populations va croire que cette corrélation est due au fait que le marqueur est lié à un QTL. Une autre façon d'expliquer ce phénomène a également été donnée en 10.1.2.

Sous l'hypothèse (H_0) $(u_1 = \ldots = u_r = 0)$ du modèle (VM_n), la statistique $\hat{T}_2$ tend en revanche vers une loi normale de moyenne 0. Ce résultat a été prouvé parAbecassis et al. (2000), qui en ont conclu que cette statistique était valide pour tester la liaison entre le marqueur et le QTL quelle que soit la structure de la population. En effet la région où l'on accepte (H_0), à savoir $[-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]$, est effectivement centrée en 0. Mais pour que le test soit vraiment de niveau α, il faut aussi que la variance soit inférieure ou égale à 1. Abecassis et al. (2000) supposent implicitement que c'est le cas mais n'ont pas étudié de

manière théorique la variance du test. Le résultat de la proposition 2 est donc intéressant car il montre que cette variance n'est en général pas égale à 1. Elle est en fait souvent supérieure.

Pour le montrer supposons, comme Abecassis et al. (2000), que

$$\sum_{i=1}^{r} \mu_i = 0$$

Sous l'hypothèse (H_0) du modèle (VM_n), on a alors

$$\mathrm{Var}(\hat{T}_2) \geq 1 \;\Leftrightarrow\; \sum_{i=1}^{r} \pi_i(1-\pi_i)\mu_i^2 \geq \sum_{i=1}^{r} \pi_i(1-\pi_i)l$$

avec

$$l = \frac{1}{r}\sum_{i=1}^{r} \mu_i^2 - \frac{(\frac{1}{r}\sum_{i=1}^{r} \mu_i(2\pi_i - 1))^2}{\frac{1}{r}\sum_{i=1}^{r}(2\pi_i-1)^2 + \frac{1}{r}\sum_{i=1}^{r}\pi_i(1-\pi_i) - \frac{1}{r^2}(\sum_{i=1}^{r}(2\pi_i-1))^2}$$

Or d'après le théorème de Cauchy-Schwarz

$$\frac{1}{r}\sum_{i=1}^{r}(2\pi_i-1)^2 - \frac{1}{r^2}(\sum_{i=1}^{r}(2\pi_i-1))^2 \geq 0$$

et donc

$$l \leq \frac{1}{r}\sum_{i=1}^{r} \mu_i^2$$

D'autre part, pour certaines configurations de π_i, il est clair qu'on peut avoir

$$\sum_{i=1}^{r} \pi_i(1-\pi_i)\mu_i^2 \geq \sum_{i=1}^{r} \pi_i(1-\pi_i)(\frac{1}{r}\sum_{i=1}^{r} \mu_i^2)$$

Dans ces cas là $\mathrm{Var}(\hat{T}_2) \geq 1$ et le test basé sur $\hat{T}_2$ est donc libéral. La correction de cette variance nécessiterait de connaître la structure, ce qui n'est pas le cas. Ce résultat va à l'encontre de l'idée reçue selon laquelle les modèles de type TDT ne présentent pas d'excès de faux positifs.

Sous l'hypothèse alternative, la variance de $\hat{T}_2$ ne change pas et cette erreur peut nuire à la puissance du test. Mais il y a également dans ce cas un problème sur la moyenne de $\hat{T}_2$, qui vaut

$$\frac{\sum_{i=1}^{r} u_i\pi_i(1-\pi_i)}{\sqrt{(\sigma^2+l)\sum_{i=1}^{r}\pi_i(1-\pi_i)}}$$

La surestimation de σ^2 et le fait que les effets u_i puissent se compenser d'une population à l'autre tendent à diminuer la moyenne du test, et donc sa puissance. Pour mieux comprendre l'influence des différents paramètres, nous considérons dans la suite quelques cas particuliers. Nous supposons toujours que

$$\sum_{i=1}^{r} \mu_i = 0$$

Une seule population

Prenons d'abord le cas extrême d'une seule population, avec les paramètres $\mu = 0$, u et π. Pour la statistique de Fisher, on trouve

$$\hat{F} \xrightarrow{\mathcal{L}} \chi^2_1(\frac{2}{\sigma^2}u^2\pi(1-\pi))$$

Pour le TDT, on a $l = 0$ et

$$\hat{T}_2 \xrightarrow{\mathcal{L}} \mathcal{L}(\frac{u}{\sigma}\sqrt{\pi(1-\pi)}, 1)$$

et donc

$$\hat{T}_2^2 \xrightarrow{\mathcal{L}} \chi^2_1(\frac{u^2}{\sigma^2}\pi(1-\pi))$$

Dans ce cas le TDT n'a pas de problème sous (H_0), et son comportement sous (VM_n) est correct. Cependant sa puissance n'est pas optimale (la moyenne de $\hat{T}_2^2$ est deux fois plus faible que celle de $\hat{F}$ car $\hat{T}_2$ utilise les variables explicatives $P_{i,j}$ qui ne sont pas nécessaires.

Fréquences homogènes entre populations

Supposons maintenant qu'il y a bien r populations distinctes, mais que les fréquences dans chaque population sont les mêmes : $\pi_1 = \ldots = \pi_r = \pi$. On trouve dans ce cas que :

$$\hat{F} \xrightarrow{\mathcal{L}} \frac{1}{r}\chi^2_r(\frac{2}{\sigma^2}\pi(1-\pi)\sum_{i=1}^{r} u_i^2)$$

Pour le TDT on a cette fois $l = \frac{1}{r}\sum_{i=1}^{r}\mu_i^2$ et

$$\hat{T}_2 \xrightarrow{\mathcal{L}} \mathcal{L}(\sqrt{\frac{\pi(1-\pi)}{r(\sigma^2+l)}}\sum_{i=1}^{r} u_i, 1)$$

c'est à dire

$$\hat{T}_2^2 \xrightarrow{\mathcal{L}} \chi^2_1(\frac{\pi(1-\pi)}{r(\sigma^2+l)}(\sum_{i=1}^{r} u_i)^2)$$

Si les fréquences des allèles au marqueur sont relativement homogènes entre populations, le TDT n'a donc pas de problème sous (H_0). Mais la puissance peut être fortement diminuée selon les valeurs des u_i et des μ_i.

Moyenne des phénotypes constante entre populations

Si les fréquences des allèles du marqueur sont inhomogènes, mais que la moyenne des phénotypes est la même dans chaque race ($\mu_1 = \ldots = \mu_r = 0$), on a toujours $l = 0$ pour le TDT, et

$$\hat{T}_2^2 \xrightarrow{\mathcal{L}} \chi_1^2\left(\frac{\left(\sum_{i=1}^r u_i \pi_i (1-\pi_i)\right)^2}{\sigma^2 \sum_{i=1}^r \pi_i (1-\pi_i)}\right)$$

Il n'y a donc toujours pas de problème sous $(M0)$, mais la puissance dépend cette fois d'une combinaison entre les u_i et les π_i dans chaque population. A titre de comparaison, on a

$$\hat{F} \xrightarrow{\mathcal{L}} \frac{1}{r}\chi_r^2\left(\frac{2}{\sigma^2}\sum_{i=1}^r u_i^2 \pi_i (1-\pi_i)\right)$$

si la structure est connue.

Effet additif commun entre populations

En général, les fréquences π_i et les moyennes μ_i sont différentes d'une population à l'autre. Le terme l a donc son expression la plus générale, et $\hat{T}_2$ n'est pas de variance 1. En revanche, l'hypothèse $u_1 = \ldots = u_r = u$ est assez commune : cela revient à supposer que le marqueur est exactement le QTL, et qu'il n'y a pas d'interaction génotype-environnement. Dans ce cas

$$\mathbb{E}[\hat{T}_2] = u\sqrt{\frac{\sum_{i=1}^r \pi_i (1-\pi_i)}{\sigma^2 + l}}$$

et la perte de puissance due à la structure ne provient que de la surestimation d'amplitude l de la variance résiduelle σ^2.

10.2.5 Résultats numériques

Pour avoir une idée plus précise de l'erreur de première espèce et de la puissance du TDT, nous l'avons calculé pour différentes valeurs des paramètres du modèle. Nous avons comparé les résultats de puissance à ceux obtenus par le test de Fisher. Pour obtenir des résultats à échelle finie, nous nous sommes placés dans le modèle (VM) pour un n fixé, et avons approché les lois des tests sous ce modèle par celles obtenues sous le modèle (VM_n) avec $u_i = \sqrt{n}a_i$. Conformément aux zones de rejet qui ont été définies en 10.2.2,

on calcule alors la puissance du test de Fisher sous (VM) par

$$1 - f(q_{1-\frac{\alpha}{2}}) + f(q_{\frac{\alpha}{2}})$$

où f est la fonction de répartition de la loi $\chi^2_r(\frac{2}{\sigma^2}\sum_{i=1}^r na_i^2\pi_i(1-\pi_i))$. La puissance du TDT sous (VM) est quant à elle évaluée par

$$1 - \varphi(z_{1-\frac{\alpha}{2}}) + \varphi(-z_{1-\frac{\alpha}{2}})$$

où φ est la distribution de la loi normale donnée par la proposition 2 avec $u_i = \sqrt{n}a_i$. Pour obtenir l'erreur de première espèce de ce test, on prend la loi normale de même variance et de moyenne 0 dans la formule ci-dessus.

Nous avons évalué les puissances des deux tests, et l'erreur de première espèce du TDT, pour différentes configurations des paramètres a_i, π_i et μ_i. Pour ces calculs, nous avons pris $r = 15$ et $n = 30$, ce qui était le cas dans le projet GemQual, et $\sigma^2 = 1$. Les coefficients μ_i dans chaque population ont été simulés à partir d'une loi $\mathcal{N}(0, v_\mu)$. Les coefficients a_i ont été simulés à partir d'une loi $\mathcal{N}(m_a, v_a)$. Enfin les fréquences π_i, qui doivent être entre 0 et 1, ont été tirés dans une loi Beta. Les deux coefficients c_1 et c_2 de cette loi ont été fixés de manière à avoir

$$\mathbb{E}[\Pi_i] = m_\pi$$

et

$$\mathrm{Var}(\Pi_i) = \frac{m_\pi(1 - m_{\pi_i})}{\lambda + 1}$$

Afin de mesurer l'influence sur la puissance de la valeur moyenne et de la variance inter populations des paramètres a_i, π_i et μ_i, nous avons fait varier m_a, v_a, m_π, λ et m_μ. Pour chaque configuration de ces cinq méta-paramètres, nous avons simulé 100 fois les vecteurs a, π et μ et calculé la moyenne des puissances (ou erreurs de première espèce) obtenues sur ces 100 configurations. Quelques résultats sont présentés ci-dessous.

Erreur de première espèce pour $\hat{T}_2$

Pour calculer l'erreur de première espèce commise avec le TDT, nous avons généré les vecteurs π et μ pour $m_\pi = 0.5$ ou 0.3, λ entre 2 et 8 et v_μ entre 1 et 5. Pour un seuil nominal $\alpha = 0.05$, l'erreur moyenne observée sur 100 populations était toujours entre 0.05 et 0.06. Le test est donc bien libéral, mai l'erreur est généralement faible, même pour des fréquences et des moyennes phénotypiques fortement inhomogènes entre populations. Ceci

explique peut-être que le problème de variance de $\hat{T}_2$ n'ait jamais été souligné auparavant.

Puissance des deux tests

	v_a					
	$m_a^2/4$		m_a^2		$4m_a^2$	
m_a	$\hat{T}_2$	$\hat{F}$	$\hat{T}_2$	$\hat{F}$	$\hat{T}_2$	$\hat{F}$
0.1	0.12	0.08	0.11	0.12	0.13	0.33
0.2	0.29	0.32	0.30	0.55	0.33	0.93
0.5	0.91	1.00	0.88	1.00	0.81	1.00

TAB. 10.1 – Puissances moyennes au niveau $\alpha = 5\%$ du TDT et du test de Fisher obtenues sur 100 populations suivant le modèle (VM). Les populations ont été simulées pour $r = 30$, $n = 15$, $v_\mu = 1$, $m_\pi = 0.5$, $\lambda = 4$ et différentes valeurs de m_a et v_a.

Les puissances du TDT et du test de Fisher sont comparées dans la table 10.2.5 pour différents types de distribution des a_i dans les $r = 15$ populations. Nous étudions les cas ou la moyenne des a_i dans l'ensemble des populations vaut 0.1, 0.2 et 0.5. Pour chacun de ces cas, nous examinons trois différents niveaux de variabilité entre populations : $\sqrt{v_a} = a/2$, $\sqrt{v_a} = a$ et $\sqrt{v_a} = 2a$. D'autre part, les fréquences π_i dans chaque population sont toujours tirées dans une loi Beta de moyenne 0.5 et de variance 0.05 (ce qui correspond à $\lambda = 4$). On prend enfin $v_\mu = 1$, ce qui signifie qu'en dehors de l'effet du marqueur la variance des phénotypes inter-populations est égale à la variance des phénotypes intra-populations ($\sigma^2 = 1$). Pour un effet génétique de moyenne faible ($m_a = 0.1$), la puissance des deux tests est mauvaise. Si la variance v_a des effets génétiques est petite, le TDT semble aussi performant que le test de Fisher, mais la comparaison n'est pas très juste car le TDT ne respecte pas le niveau α, et l'impact de cette erreur est d'autant plus grand qu'on s'approche de (H_0). De manière générale le TDT est naturellement moins puissant, mais c'est surtout si les effets génétiques sont très variable de population en population que la différence est énorme. Pour des effets relativement homogènes entre populations, le TDT donne des résultats corrects.

v_μ	0.5	1	2	3
Puissance	0.96	0.91	0.79	0.61

TAB. 10.2 – Puissances moyennes au niveau $\alpha = 5\%$ du TDT obtenues sur 100 populations suivant le modèle (VM). Les populations ont été simulées pour $r = 30$, $n = 15$, $m_a = 0.5$, $v_a = m_a^2/4 = 0.06$, $m_\pi = 0.5$, $\lambda = 4$ et différentes valeurs de v_μ.

La puissance du TDT est également affectée par l'hétérogénéité des μ_i, ce qui est illustré par la table 10.2.5. Nous avons repris dans cette table une des configurations de la table 10.2.5 les plus favorables au TDT ($m_a = 0.5$ et $v_a = m_a^2/4$) et avons fait varier v_μ. Les résultats montrent que la puissance diminue assez vite quand la variance des μ_i augmente, alors que cela n'a pas d'influence sur le test de Fisher. Pour le projet GemQual, et plus généralement pour la recherche de QTL en agriculture ou élevage, c'est un problème à prendre en compte. En effet, au sein de chaque population, les individus sont souvent fortement sélectionnés par rapport à certains caractères, ce qui tend à augmenter les différence entre les races vis à vis de ces caractères.

Rappelons enfin que les fréquences des allèles ont également une influence qui peut être importante sur la puissance des tests, pour le TDT comme pour le test de Fisher. Nous avons pris ici des fréquences relativement homogènes et proches de 0.5, ce qui est le cas idéal pour la détection des QTL. Pour des fréquences plus proches de 0 ou 1 les résultats seraient moins bons.

10.2.6 Preuve de la proposition 2

Revenons à présent sur la démonstration de la proposition 2, qui donne la limite en loi des statistiques $\hat{F}$, $\hat{T}_1$ et $\hat{T}_2$ sous le modèle (VM_n) quand n tend vers l'infini. La démarche utilisée est à peu prés la même pour les 3 statistiques : nous essayons d'obtenir, à partir des définitions données en 10.2.2, une expression assez simple de la statistique en fonction des variables $Z_{i,j}$, $M_{i,j}$, $P_{i,j}$ (s'il y a lieu) et $\epsilon_{i,j}$. Nous remplaçons ensuite $Z_{i,j}$ par son expression dans (VM_n), à savoir

$$Z_{i,j} = \mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i) + \epsilon_{i,j}$$

Finalement nous appliquons la loi des grands nombres ou le théorème central limite en utilisant les moments des $M_{i,j}$, $P_{i,j}$ et $\epsilon_{i,j}$ donnés en 10.2.1.

Pour la limite de $\hat{F}$, on pourrait en fait déduire le résultat de travaux classiques sur le modèle linéaire, dans la mesure où $\hat{F}$ est la statistique de Fisher utilisée naturellement pour rejeter (H_0) dans le modèle (VM) et qu'il s'agit simplement ici de donner sa loi sous une hypothèse alternative locale. Le résultat n'a donc rien de nouveau en soi, mais nous avons toutefois jugé intéressant d'en donner une preuve directe. En revanche le cas de $\hat{T}_1$ et $\hat{T}_2$ est plus singulier. Puisque ces statistiques proviennent de modèles d'analyse différents de (VM), elles n'ont aucune propriété particulière sous (VM). Leur loi doit donc être étudiée "à la main".

Limite de la statistique de test $\hat{F}$

Rappelons que

$$\hat{F} = \frac{(SCR_0 - SCR_1)/r}{SCR_1/(rn-r)}$$

avec

$$SCR_1 = ||Z - P_X Z||^2$$

et

$$SCR_0 = ||Z - P_{\underline{X}} Z||^2$$

Tout d'abord il est clair que $SCR_1/(rn-r) \xrightarrow{p} \sigma^2$. En effet on sait que $Z - P_X Z = \epsilon - P_X \epsilon$. D'autre part les $\epsilon_{i,j}$ sont de moyenne nulle et indépendants des $M_{i,j}$ donc ϵ est asympotiquement orthogonal à E_X. Par conséquent

$$\begin{aligned} SCR_1/(rn-r) &= \frac{1}{rn-r}||\epsilon||^2 \\ &= \frac{n}{rn-r}\sum_{i=1}^{r}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{i,j}^2 \\ &\xrightarrow{p} \frac{1}{r}\sum_{i=1}^{r}\sigma^2 = \sigma^2 \end{aligned}$$

D'autre part, on a

$$\begin{aligned} SCR_1 &= ||Z - P_{\underline{X}} Z||^2 \\ &= ||Z - P_X Z||^2 + ||P_X Z - P_{\underline{X}} Z||^2 \end{aligned}$$

puisque $E_{\underline{X}}$ est inclus dans E_X. Par conséquent $SCR_1 - SCR_0 = ||P_X Z - P_{\underline{X}} Z||^2$. On introduit la notation $X = (X_{\mu_1}, \ldots, X_{\mu_r}, X_{u_1}, \ldots, X_{u_r})$, où les X_{μ_i} et X_{u_i} sont des vecteurs de taille $rn \times 1$. Ces vecteurs sont tous orthogonaux entre eux. Comme $\underline{X} = (X_{\mu_1}, \ldots, X_{\mu_r})$, on obtient donc

$$\begin{aligned} SCR_1 - SCR_0 &= ||\sum_{i=1}^{r} P_{X_{a_i}} Z||^2 \\ &= ||\sum_{i=1}^{r} \frac{\sum_{j=1}^{n} Z_{i,j}(M_{i,j} - \overline{M}_i)}{\sum_{j=1}^{n}(M_{i,j} - \overline{M}_i)^2} X_{a_i}||^2 \\ &= \sum_{i=1}^{r} \frac{(\sum_{j=1}^{n} Z_{i,j}(M_{i,j} - \overline{M}_i))^2}{\sum_{j=1}^{n}(M_{i,j} - \overline{M}_i)^2} \end{aligned}$$

Or on sait que

$$\frac{1}{n}\sum_{j=1}^{n}(M_{i,j}-\overline{M}_i)^2 \xrightarrow{p} \mathrm{Var}(M_{i,1})$$

et que

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{j=1}^{n} Z_{i,j}(M_{i,j}-\overline{M}_i) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\mu_i+\frac{u_i}{\sqrt{n}}(M_{i,j}-\overline{M}_i)+\epsilon_{i,j})(M_{i,j}-\overline{M}_i)\\ &= \frac{u_i}{n}\sum_{j=1}^{n}(M_{i,j}-\overline{M}_i)^2+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(M_{i,j}-\overline{M}_i)\epsilon_{i,j}\end{aligned}$$

Le premier terme tend en probabilités vers $u_i\mathrm{Var}(M_{i,1})$. Le second tend en distribution vers une normale $\mathcal{N}(0,\sigma^2\mathrm{Var}(M_{i,1}))$. En effet

$$\mathbb{E}[\epsilon_{i,1}(M_{i,1}-\mathbb{E}[M_{i,1}])]=0$$

et

$$\mathbb{E}[\epsilon_{i,1}^2(M_{i,1}-\mathbb{E}[M_{i,1}])^2]=\sigma^2\mathrm{Var}(M_{i,1})$$

Sachant que $\mathrm{Var}(M_{i,1})=2\pi_i(1-\pi_i)$ et que les variables sont indépendantes d'une population à une autre, le résultat est ensuite immédiat.

Limite de la statistique de test $\hat{T}_1$

On a

$$\hat{T}_1=\frac{\hat{a}_0}{\sqrt{\hat{\sigma}^2}}\sqrt{\sum_{i=1}^{r}\sum_{j=1}^{n}(M_{i,j}-\overline{M})^2}$$

avec

$$\hat{a}_0=\frac{\sum_{i=1}^{r}\sum_{j=1}^{n}Z_{i,j}(M_{i,j}-\overline{M})}{\sum_{i=1}^{r}\sum_{j=1}^{n}(M_{i,j}-\overline{M})^2}$$

L'idée est ici de montrer que $\hat{a}_0$ ne tend pas vers 0. En effet on a d'autre part

$$\sqrt{\sum_{i=1}^{r}\sum_{j=1}^{n}(M_{i,j}-\overline{M})^2}\xrightarrow{p}+\infty$$

et on voit bien, sans avoir à rentrer dans les détails, que $\hat{\sigma}^2\xrightarrow{p} l>0$.

Or dans le modèle (VM_n)

$$Z_{i,j}(M_{i,j} - \overline{M}) = \mu_i(M_{i,j} - \overline{M}) + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i)(M_{i,j} - \overline{M}) + \epsilon_{i,j}(M_{i,j} - \overline{M})$$

Comme $\epsilon_{i,j}$ est indépendant de $M_{i,j}$, on montre facilement par la loi des grands nombres que quand n tend vers l'infini

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n} Z_{i,j}(M_{i,j} - \overline{M}) &\xrightarrow{p} \sum_{i=1}^{r}\mu_i\mathbb{E}[M_{i,1}] - \frac{1}{r}(\sum_{i=1}^{r}\mu_i)(\sum_{i=1}^{r}\mathbb{E}[M_{i,j}]) \\ &= \sum_{i=1}^{r}\mu_i(2\pi_i - 1) - \frac{1}{r}(\sum_{i=1}^{r}\mu_i)(\sum_{i=1}^{r}(2\pi_i - 1))\end{aligned}$$

On a de même

$$\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n}(M_{i,j} - \overline{M})^2 \xrightarrow{p} v > 0$$

D'où le résultat.

Limite de la statistique de test $\hat{T}_2$

Nous avons défini

$$\hat{T}_2 = \frac{\hat{\beta}_w}{\sqrt{\hat{\sigma}^2}\sqrt{e_3({}^t\widetilde{X}\widetilde{X})^{-1}({}^te_3)}}$$

où

$$\begin{aligned}\hat{\beta}_w &= e_3({}^t\widetilde{X}\widetilde{X})^{-1}({}^t\widetilde{X}Z) \\ \hat{\sigma}^2 &= \frac{||Z - P_{\widetilde{X}}Z||^2}{rn - 3}\end{aligned}$$

On remarque tout d'abord que $\frac{1}{n}({}^t\widetilde{X}\widetilde{X})$ est asymptotiquement diagonale. En effet X_{β_w} est orthogonal à X_{μ_0} et à X_{β_b} quand $n \to \infty$, puisque

$$\frac{1}{n}\,{}^tX_{\beta_w}X_{\mu_0} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n}(M_{i,j} - P_{i,j}) \xrightarrow{p} \sum_{i=1}^{r}\mathbb{E}[M_{i,1} - P_{i,1}]$$

et

$$\begin{aligned}\frac{1}{n}\,{}^tX_{\beta_w}X_{\beta_b} &= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n}(M_{i,j}-P_{i,j})(P_{i,j}-\overline{P}) \\ &\xrightarrow{p} \sum_{i=1}^{r}\mathbb{E}[(M_{i,1}-P_{i,1})P_{i,1}] - \frac{1}{r}(\sum_{i=1}^{r}\mathbb{E}[P_{i,1}])(\sum_{i=1}^{r}\mathbb{E}[M_{i,1}-P_{i,1}])\end{aligned}$$

Or les formules présentées en 10.2.1 permettent de déduire que $\mathbb{E}[M_{i,1}-P_{i,1}]=0$ et $\mathbb{E}[(M_{i,1}-P_{i,1})P_{i,1}]=0$. D'autre part on sait que X_{μ_0} est toujours orthogonal à X_{β_b}.
Quand $n\to\infty$, on peut donc écrire

$$\hat{T}_2 = \frac{e_3(\frac{1}{n}\,{}^t\widetilde{X}\widetilde{X})^{-1}(\frac{1}{\sqrt{n}}\,{}^t\widetilde{X}Z)}{\sqrt{\hat{\sigma}^2}\sqrt{e_3(\frac{1}{n}\,{}^t\widetilde{X}\widetilde{X})^{-1}({}^te_3)}} = \frac{\frac{1}{\sqrt{n}}\,{}^tX_{\beta_w}Z}{\sqrt{\hat{\sigma}^2}\sqrt{\frac{1}{n}\,{}^tX_{\beta_w}X_{\beta_w}}} + o_p(1)$$

où la notation $o_p(1)$ désigne une variable aléatoire qui tend en probabilités vers 0. Or on a

$$\frac{1}{n}\,{}^tX_{\beta_w}X_{\beta_w} = \sum_{i=1}^{r}\frac{1}{n}\sum_{j=1}^{n}(M_{i,j}-P_{i,j})^2 \xrightarrow{p} \sum_{i=1}^{r}\mathbb{E}[(M_{i,1}-P_{i,1})^2] = \sum_{i=1}^{r}\pi_i(1-\pi_i)$$

Il ne reste donc plus qu'à étudier la limite de $\frac{1}{\sqrt{n}}\,{}^tX_{\beta_w}Z$ et de $\hat{\sigma}^2$. Ces limites sont présentées sous la forme de deux lemmes.

Lemma 1 *Si les variables $M_{i,j}$ et $P_{i,j}$ suivent la loi décrite en 10.2.1 et si la loi des variables $Z_{i,j}$ est déterminée par le modèle $(M1_n)$, alors :*

$$\frac{1}{\sqrt{n}}\,{}^tX_{\beta_w}Z \xrightarrow{\mathcal{L}} \mathcal{N}(\sum_{i=1}^{r}u_i\pi_i(1-\pi_i), \sum_{i=1}^{r}\pi_i(1-\pi_i)(\mu_i^2+\sigma^2))$$

Preuve : On sait que

$$\frac{1}{\sqrt{n}}\,{}^tX_{\beta_w}Z = \sum_{i=1}^{r}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(M_{i,j}-P_{i,j})Z_{i,j}$$

Or sous (VM_n) on peut écrire

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(M_{i,j}-P_{i,j})Z_{i,j} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(M_{i,j}-P_{i,j})(\mu_i+\epsilon_{i,j}) + \frac{u_i}{n}\sum_{j=1}^{n}(M_{i,j}-P_{i,j})(M_{i,j}-\overline{M}_i)$$

Pour trouver la limite du second terme de cette expression, on calcule

$$\begin{aligned}\mathbb{E}[(M_{i,1}-P_{i,1})(M_{i,1}-\mathbb{E}[M_{i,1}])] &= \mathbb{E}[(M_{i,1}-\mathbb{E}[P_{i,1}])(M_{i,1}-\mathbb{E}[M_{i,1}])] \\ &- \mathbb{E}[(P_{i,1}-\mathbb{E}[P_{i,1}])(M_{i,1}-\mathbb{E}[M_{i,1}])] \\ &= \mathrm{Var}(M_{i,1})-\mathrm{Cov}(P_{i,1},M_{i,1})\end{aligned}$$

car $\mathbb{E}[P_{i,1}]=\mathbb{E}[M_{i,1}]$. En utilisant les expressions données en 10.2.1, on aboutit finalement à

$$\frac{u_i}{n}\sum_{j=1}^{n}(M_{i,j}-P_{i,j})(M_{i,j}-\overline{M}_i)\xrightarrow{p}u_i\pi_i(1-\pi_i)$$

Pour la limite de l'autre terme, posons $U_{i,j}=(M_{i,j}-P_{i,j})(\mu_i+\epsilon_{i,j})$. On a

$$\mathbb{E}[U_{i,j}]=\mathbb{E}[M_{i,j}-P_{i,j}]\mathbb{E}[\mu_i+\epsilon_{i,j}]=0$$

et

$$\mathrm{Var}(U_{i,j})=\mathbb{E}[(M_{i,j}-P_{i,j})^2]\mathbb{E}[(\mu_i+\epsilon_{i,j})^2]=\pi_i(1-\pi_i)(\mu_i^2+\sigma^2)$$

On en déduit que

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(M_{i,j}-P_{i,j})(\mu+\epsilon_{i,j})\xrightarrow{\mathcal{L}}\mathcal{N}(0,\pi_i(1-\pi_i)(\mu_i^2+\sigma^2))$$

D'où le résultat.

Lemma 2 *Si les variables $M_{i,j}$ et $P_{i,j}$ suivent la loi décrite en 10.2.1 et si la loi des variables $Z_{i,j}$ est déterminée par le modèle $(M1_n)$, alors :*

$$\hat{\sigma}^2\xrightarrow{p}\sigma^2+l1-l2-l3$$

où

$$\begin{aligned}l1 &= \frac{1}{r}\sum_{i=1}^{r}\mu_i^2 \\ l2 &= \frac{1}{r^2}(\sum_{i=1}^{r}\mu_i)^2 \\ l3 &= \frac{(\frac{1}{r}\sum_{i=1}^{r}(\mu_i(2\pi_i-1))-\frac{1}{r^2}(\sum_{i=1}^{r}\mu_i)(\sum_{i=1}^{r}(2\pi_i-1)))^2}{\frac{1}{r}\sum_{i=1}^{r}(2\pi_i-1)^2+\frac{1}{r}\sum_{i=1}^{r}\pi_i(1-\pi_i)-\frac{1}{r^2}(\sum_{i=1}^{r}(2\pi_i-1))^2}\end{aligned}$$

Preuve : On utilise la décomposition $Z = X\theta + \epsilon$ issue du modèle $(M1_n)$, qui donne :

$$||Z - P_{\widetilde{X}}Z||^2 = ||Z||^2 - ||P_{\widetilde{X}}Z||^2 = ||X\theta + \epsilon||^2 - ||P_{\widetilde{X}}(X\theta + \epsilon)||^2$$

D'après les hypothèses du modèle $(M1_n)$, les $\epsilon_{i,j}$ sont de moyenne nulle et indépendants des $M_{i,j}$ et des $P_{i,j}$ donc ϵ est asympotiquement orthogonal aux espaces E_X et $E_{\widetilde{X}}$. Par conséquent on a

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{N_s - 3}||Z - P_{\widetilde{X}}Z||^2 \\
&= \frac{1}{N_s}(||X\theta||^2 + ||\epsilon||^2 - ||P_{\widetilde{X}}(X\theta)||^2) + o_p(1) \\
&= \sigma^2 + \frac{1}{N_s}(< X\theta, X\theta > - < P_{\widetilde{X}}(X\theta), P_{\widetilde{X}}(X\theta) >) + o_p(1) \\
&= \sigma^2 + \frac{1}{N_s}(< X\theta, X\theta > - < P_{\widetilde{X}}(X\theta), X\theta >) + o_p(1)
\end{aligned}$$

On trouve d'abord que

$$\begin{aligned}
S1 &= \frac{1}{N_s} < X\theta, X\theta > \\
&= \frac{1}{N_s}\sum_{i=1}^{r}\sum_{j=1}^{n}(\mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i))^2 \\
&\xrightarrow{p} \frac{1}{r}\sum_{i=1}^{r}\mu_i^2
\end{aligned}$$

puisque les termes en $\frac{a_i}{\sqrt{n}}$ et $\frac{a_i^2}{n}$ sont négligeables. On obtient ainsi le terme $l1$. D'autre part, on a vu que les vecteurs X_{μ_0}, X_{β_b} et X_{β_w} étaient asymptotiquement orthogonaux, donc on a

$$\frac{1}{N_s} < P_{\widetilde{X}}(X\theta), X\theta >= S2 + S3 + S4 + o_p(1)$$

avec

$$\begin{aligned}
S2 &= \frac{1}{N_s} < P_{X_{\mu_0}}(X\theta), X\theta > \\
S3 &= \frac{1}{N_s} < P_{X_{\beta_b}}(X\theta), X\theta > \\
S4 &= \frac{1}{N_s} < P_{X_{\beta_w}}(X\theta), X\theta >
\end{aligned}$$

Ces trois termes tendent respectivement vers $l2$, $l3$ et 0. En effet :

$$\begin{aligned} S2 &= \frac{1}{N_s} < (\frac{1}{N_s}\sum_{i=1}^{r}\sum_{j=1}^{n}(\mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i)))X_{\mu_0}, X\theta > \\ &= \frac{1}{N_s^2}(\sum_{i=1}^{r}\sum_{j=1}^{n}(\mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i)))^2 \\ &\xrightarrow{p} \frac{1}{r^2}(\sum_{i=1}^{r}\mu_i)^2 \end{aligned}$$

$$\begin{aligned} S3 &= \frac{1}{N_s} < \frac{\sum_{i=1}^{r}\sum_{j=1}^{n}(\mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i))(P_{i,j} - \overline{P})}{\sum_{i=1}^{r}\sum_{j=1}^{n}(P_{i,j} - \overline{P})^2} X_{\beta_b}, X\theta > \\ &= \frac{(\frac{1}{N_s^2}\sum_{i=1}^{r}\sum_{j=1}^{n}(\mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i))(P_{i,j} - \overline{P}))^2}{\frac{1}{N_s}\sum_{i=1}^{r}\sum_{j=1}^{n}(P_{i,j} - \overline{P})^2} \\ &= \frac{S31^2}{S32} \end{aligned}$$

avec

$$\begin{aligned} S31 &\xrightarrow{p} \frac{1}{r}\sum_{i=1}^{r}(\mu_i\mathbb{E}[P_{i,1}]) - \frac{1}{r^2}(\sum_{i=1}^{r}\mu_i)(\sum_{i=1}^{r}\mathbb{E}[P_{i,1}]) \\ &= \frac{1}{r}\sum_{i=1}^{r}(\mu_i(2\pi_i - 1)) - \frac{1}{r^2}(\sum_{i=1}^{r}\mu_i)(\sum_{i=1}^{r}(2\pi_i - 1)) \end{aligned}$$

et

$$\begin{aligned} S32 &\xrightarrow{p} \frac{1}{r}\sum_{i=1}^{r}\mathbb{E}[P_{i,1}^2] - \frac{1}{r^2}(\sum_{i=1}^{r}\mathbb{E}[P_{i,1}])^2 \\ &= \frac{1}{r}\sum_{i=1}^{r}(2\pi_i - 1)^2 + \frac{1}{r}\sum_{i=1}^{r}\pi_i(1 - \pi_i) - \frac{1}{r^2}(\sum_{i=1}^{r}(2\pi_i - 1))^2 \end{aligned}$$

Enfin

$$\begin{aligned} S4 &= \frac{1}{N_s} < \frac{\sum_{i=1}^{r}\sum_{j=1}^{n}(\mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i))(M_{i,j} - P_{i,j})}{\sum_{i=1}^{r}\sum_{j=1}^{n}(M_{i,j} - P_{i,j})^2} X_{\beta_w}, X\theta > \\ &= \frac{\frac{1}{N_s^2}(\sum_{i=1}^{r}\sum_{j=1}^{n}(\mu_i + \frac{u_i}{\sqrt{n}}(M_{i,j} - \overline{M}_i))(M_{i,j} - P_{i,j}))^2}{\frac{1}{N_s}\sum_{i=1}^{r}\sum_{j=1}^{n}(M_{i,j} - P_{i,j})^2} \\ &\to 0 \end{aligned}$$

car $\mathbb{E}[M_{i,1} - P_{i,1}] = 0$. D'où le résultat.

10.2.7 Conclusions

Nous avons dérivé dans ce chapitre la loi asymptotique de trois statistiques d'association dans un modèle de population structurée. La première statistique $\hat{F}$, qui nous servait de référence, supposait connue la structure de la population, alors que les deux autres, $\hat{T}_1$ et $\hat{T}_2$, partaient du principe qu'elle était inconnue. Comme l'avaient déjà souligné plusieurs auteurs, nous avons pu vérifié que le test $\hat{T}_1$ n'est pas fiable dans le cas d'une population structurée. Par la suite nous nous sommes plutôt concentrés sur le test $\hat{T}_2$, qui est un test de type TDT proposé par (Abecassis et al., 2000).

Les résultats obtenus pour $\hat{T}_2$ sont intéressants. Bien que ce test soit de moyenne nulle sous l'hypothèse (H_0) d'indépendance entre le marqueur et le caractère, nos résultats montrent qu'il est libéral car sa variance est généralement supérieure à 1. Nous avons cependant constaté sur des populations simulées que l'erreur de première espèce n'était que peu supérieure au niveau du test. Le fait de connaître la distribution sous l'hypothèse alternative nous a également permis d'étudier la puissance de ce test. Sur ce point, les résultats obtenus sont conformes à ce qu'on pouvait attendre. Ils indiquent que pour des populations très hétérogènes (grands écarts entre les μ_i et les a_i d'une population à l'autre), la perte de puissance du TDT par rapport à $\hat{F}$ peut être très importante. Mais pour des populations relativement homogènes, le TDT reste compétitif. L'intérêt de notre travail est qu'il permet de quantifier la perte de puissance due à la structure en fonction du nombre de sous populations, des fréquences alléliques dans les populations ...

Pour une population qui se présente sous forme de sous-populations distinctes et connues, comme dans le projet GeMqual par exemple, on peut malgré tout avoir envie d'utiliser le TDT pour se prémunir d'une éventuelle structure interne aux sous-populations. Dans ce cas il peut être utile d'avoir une idée de la puissance qu'on peut espérer obtenir avec ce test. Si la structure de la population est a priori complètement inconnue, on peut toujours essayer de l'estimer, par exemple à l'aide de la méthode de Pritchard, Stephens, et Donnelly (2000). Nos résultats peuvent alors permettre de décider s'il est plus prometteur d'utiliser cette estimation de la structure et un test de Fisher, ou bien directement un TDT.

Références

Abecassis, G., Cardon, L., et Cookson, W. (2000). A general test of association for quantitaive traits in nuclear families. *Am. J. Hum. Genet.*, *66*, 279-292.

Allison, D. (1997). Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.*, *60*, 676-690.

Benjamini, Y., et Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289-300.

Cierco-Ayrolles, C., Abdallah, J., Boitard, S., Chikhi, L., Rochambeau, H. de, Tsitrone, A., et al. (2004). On linkage disequilibrium measures : Methods and applications. In (Vol. 1, p. 151-180). Research Signpost, India.

Devlin, B., et Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*, 997-1004.

Köhler, K., et Bickeböller, H. (2005). Case-control association tests correcting for population stratification. *Ann. Hum. Genet.*, *69*, 98-115.

McIntyre, L., Martin, E., Simonsen, K., et Kaplan, N. (2000). Cicrcumventing multiple testing : a multilocus monte carlo approach to testing for association. *Genetic Epidemiology*, *19*, 18-29.

Nielsen, D., et Weir, B. (1999). A classical setting for associations between markers and loci affecting quantitative traits. *Genet. Res*, *74*, 271-277.

Nyholt, D. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.*, *74*, 765-769.

Pritchard, J., Stephens, M., et Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945-959.

Pritchard, J., Stephens, M., Rosenberg, N., et Donnelly, P. (2000). Association mapping in structures population. *Am. J. Hum. Genet.*, *67*, 170-181.

Spielman, R., McGinnis, R., et Ewens, W. (1993). Transmission test for linkage disequilibrium : the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am. J. Hum. Genet.*, *52*, 506-516.

Weller, J., Song, J., Heyen, D., Lewin, H., et Ron, M. (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics*, *150*, 1699-1706.

Conclusion générale

J'ai présenté dans ce manuscrit les différents travaux de recherche que j'ai effectués dans le domaine de la cartographie de QTL par déséquilibre de liaison. Voici pour finir un résumé des principaux résultats que j'ai obtenus et des perspectives qu'ils ouvrent.

J'ai proposé dans la partie II un algorithme numérique permettant de calculer de manière approchée la densité de transition des fréquences d'haplotypes sous un modèle de diffusion à deux loci bialléliques avec recombinaison. C'est un problème classique de génétique des populations, dont la solution exacte n'est pas connue. La solution que je propose est assez générale, puisque l'algorithme peut être aussi utilisé pour des modèles incluant sélection et mutation. Cette solution est certes approchée, mais j'ai réussi à identifier les cas où elle n'était pas assez précise, et j'ai proposé des pistes d'amélioration. Celles-ci constituent un axe de recherche possible pour l'avenir.

Connaître la loi du processus de diffusion à deux loci liés a non seulement un intérêt théorique important, mais aussi des applications extrèmement nombreuses. Certaines ont été évoquées à la fin de la partie II. J'ai présenté notamment des résultats préliminaires qui montrent que ma méthode pourrait permettre d'améliorer la précision des méthodes de cartographie génétique fréquentistes, pour un temps de calcul qui resterait tout à fait raisonnable. Je m'étais cependant placé dans un cadre assez idéal : le génotype du QTL était directement observable, l'âge de la mutation et la taille efficace de la population étaient connus .... Beaucoup de travail reste à faire pour pouvoir utiliser cette méthode dans des situations concrètes, et la comparer dans ces cas à d'autres méthodes existantes.

La méthode présentée dans la partie III est quant tout à fait concrète et applicable. C'est une méthode de cartographie fine de QTL, qui présente l'avantage d'être extrèmement rapide tout en utilisant l'information d'haplotypes. Pour l'instant ces haplotypes ne sont constitués que des deux marqueurs flanquants pour la position à évaluer, mais les performances sont dèjà preque au niveau de celles obtenues par des méthodes utilisant un nombre quelconque de marqueurs. Une perspective évidente est d'étendre notre méthode pour lui permettre d'utiliser plus de marqueurs à la fois et de rester valide dans des mo-

dèles de population plus généraux que ceux considérés jusqu'à maintenant. J'ai proposé en fin de partie III des pistes pour aller dans ce sens.

Enfin, j'ai étudié dans la partie IV la loi asymptotique d'un test de type TDT quantitatif dans un modèle de population structurée assez général. Cela a permis de montrer que ce test était légèrement libéral, contrairement à ce qui souvent avancé. Mes résultats permettent en outre de calculer la puissance de ce test en fonction de la valeur des vrais paramètres de la population structurée. Cependant, je fais l'hypothèse que les individus observés sont répartis de manière égale dans toutes les sous-populations. Il serait intéressant de relaxer cette hypothèse. D'autre part le raisonnement employé dans cette partie pourrait également servir à étudier d'autres tests d'associations que ceux que j'ai considérés.

Annexes

Annexe A

Article : On linkage disequilibrium measures : methods and applications

Annexe B

Compléments de démonstration

B.1 Probabilité de transition pour le modèle de Wright-Fisher à deux loci

Soit $X(t)$ le vecteur des effectifs au temps t sous un modèle de Wright-Fisher à deux loci. Il a été dit en 2.1.2 que la loi de $X(t+1)$ sachant $X(t)$ était une multinomiale $\mathcal{M}(2N, p)$ avec

$$p_{i_1,i_2} = (1-c)\Pi_{i_1,i_2}(t) + c\Pi_{i_1,.}(t)\Pi_{.,i_2}(t) \qquad i_1 = 1,\dots,I_1,\ i_2 = 1,\dots,I_2$$

Nous montrons ici que cette loi est bien cohérente avec le processus de simulation décrit en 2.1.2 dans le cadre général.

Pour cela suivons pas à pas le déroulement du procédé de simulation. On tire d'abord deux haplotypes selon une loi $\mathcal{M}(2N, \Pi(t))$ (étape 2a). Soient U et V les variables aléatoires correspondant à ces deux tirages. Soit également H la variable représentant l'haplotype formé à l'issue du processus de simulation. On note $H = (H_1, H_2)$, où H_1 représente l'allèle au locus 1 et H_2 l'allèle au locus 2. La loi de H est décrite par les probabilités

$$p_{i_1,i_2} = \mathbb{P}(H = (i_1, i_2)),\ i_1 = 1,\dots,I_1,\ i_2 = 1,\dots,I_2$$

Pour un couple (i_1, i_2) donné, cette probabilité peut s'écrire

$$\begin{aligned} p_{i_1,i_2} &= \sum_{(m_1,m_2)} \sum_{(n_1,n_2)} \mathbb{P}(U = (m_1,m_2))\mathbb{P}(V = (n_1,n_2))\theta_{i_1,i_2:m_1,m_2|n_1,n_2} \\ &= \sum_{(m_1,m_2)} \sum_{(n_1,n_2)} \Pi_{m_1,m_2}(t)\Pi_{n_1,n_2}(t)\theta_{i_1,i_2:m_1,m_2|n_1,n_2} \end{aligned}$$

où $\theta_{i_1,i_2:m_1,m_2|n_1,n_2} = \mathbb{P}(H = (i_1,i_2) \mid U = (m_1,m_2), V = (n_1,n_2))$. A l'étape 2b de la simulation, on choisit au hasard le premier allèle de H entre m_1 et n_1. Cela se traduit par

$$\begin{aligned}\theta_{i_1,i_2:m_1,m_2|n_1,n_2} &= \frac{1}{2}\delta_{i_1,m_1}\mathbb{P}(H_2 = i_2 \mid U = (m_1,m_2), V = (n_1,n_2), V_1 = m_1) \\ &+ \frac{1}{2}\delta_{i_1,n_1}\mathbb{P}(H_2 = i_2 \mid U = (m_1,m_2), V = (n_1,n_2), V_1 = n_1)\end{aligned}$$

A l'étape 2c on choisit enfin l'allèle H_2 entre m_2 et n_2 en tenant compte de la recombinaison. On obtient donc les 2 équations

$$\mathbb{P}(H_2 = i_2 \mid U = (m_1,m_2), V = (n_1,n_2), H_1 = m_1) = (1-c)\delta_{i_2,m_2} + c\delta_{i_2,n_2}$$

et

$$\mathbb{P}(H_2 = i_2 \mid U = (m_1,m_2), V = (n_1,n_2), H_1 = n_1) = (1-c)\delta_{i_2,n_2} + c\delta_{i_2,m_2}$$

On aboutit finalement à

$$\begin{aligned}p_{i_1,i_2} &= \frac{1}{2}\sum_{(m_1,m_2)}\sum_{(n_1,n_2)}\Pi_{m_1,m_2}(t)\Pi_{n_1,n_2}(t) \\ &* \ [\delta_{i_1,m_1}((1-c)\delta_{i_2,m_2} + c\delta_{i_2,n_2}) + \delta_{i_1,n_1}((1-c)\delta_{i_2,n_2} + c\delta_{i_2,m_2})] \\ &= (1-c)\Pi_{i_1,i_2}(t) + c(\sum_{m_2}\Pi_{i_1,m_2}(t))\sum_{m_1}\Pi_{(m_1,i_2)}(t)) \\ &= (1-c)\Pi_{i_1,i_2}(t) + c\Pi_{i_1,\cdot}(t)\Pi_{\cdot,i_2}(t)\end{aligned}$$

On retrouve bien la probabilité voulue. Et comme dans la simulation on itère $2N$ fois ce processus de manière indépendante on aboutit bien à la loi multinomiale $\mathcal{M}(2N,p)$.

B.2 Développement de l'équation projective de Kolmogorov

Nous montrons ici comment passer, dans le cas de deux loci bialléliques, de la forme compacte de l'équation projective de Kolmogorov à la forme développée présentée en 5.3. Les notations sont les mêmes que dans cette section. Pour simplifier on écrit seulement f

pour $f(y,\tau)$. On a

$$\begin{aligned}
\frac{\partial f}{\partial \tau} &= -\sum_{j=1}^{3}\frac{\partial (b_j(y)f)}{\partial y_j}+\frac{1}{2}\sum_{j_1,j_2=1}^{3}\frac{\partial^2(a^2_{j_1,j_2}(y)f)}{\partial y_{j_1}\partial y_{j_2}}\\
&= -\sum_{j=1}^{3}\frac{\partial b_j(y)}{\partial y_j}f-\sum_{j=1}^{3}b_j(y)\frac{\partial f}{\partial y_j}\\
&+ \frac{1}{2}\sum_{j=1}^{3}\frac{\partial^2 a^2_{j,j}(y)}{\partial y_j^2}f+\frac{1}{2}\sum_{j=1}^{3}2\frac{\partial a^2_{j,j}(y)}{\partial y_j}\frac{\partial f}{\partial y_j}+\frac{1}{2}\sum_{j=1}^{3}a^2_{j,j}(y)\frac{\partial^2 f}{\partial y_j^2}\\
&+ \frac{1}{2}\sum_{j_1\neq j_2}\frac{\partial^2 a^2_{j_1,j_2}(y)}{\partial y_{j_1}\partial y_{j_2}}f+\frac{1}{2}\sum_{j_1\neq j_2}(\frac{\partial a^2_{j_1,j_2}(y)}{\partial y_{j_1}}\frac{\partial f}{\partial y_{j_2}}+\frac{\partial a^2_{j_1,j_2}(y)}{\partial y_{j_2}}\frac{\partial f}{\partial y_{j_1}})\\
&+ \frac{1}{2}\sum_{j_1\neq j_2}a^2_{j_1,j_2}(y)\frac{\partial^2 f}{\partial y_{j_1}\partial y_{j_2}}\\
&= \left(\frac{1}{2}\sum_{j=1}^{3}\frac{\partial^2 a^2_{j,j}(y)}{\partial y_j^2}+\frac{1}{2}\sum_{j_1\neq j_2}\frac{\partial^2 a^2_{j_1,j_2}(y)}{\partial y_{j_1}\partial y_{j_2}}-\sum_{j=1}^{3}\frac{\partial b_j(y)}{\partial y_j}\right)f\\
&+ \sum_{j_1=1}^{3}\left(\frac{\partial a^2_{j_1,j_1}(y)}{\partial y_{j_1}}+\sum_{j_2\neq j_1}\frac{\partial a^2_{j_1,j_2}(y)}{\partial y_{j_2}}-b_{j_1}(y)\right)\frac{\partial f}{\partial y_{j_1}}\\
&+ \frac{1}{2}\sum_{j=1}^{3}a^2_{j,j}(y)\frac{\partial^2 f}{\partial y_j^2}+\frac{1}{2}\sum_{j_1\neq j_2}a^2_{j_1,j_2}(y)\frac{\partial^2 f}{\partial y_{j_1}\partial y_{j_2}}\\
&= \left(\frac{1}{2}\sum_{j=1}^{3}\frac{\partial^2 a^2_{j,j}(y)}{\partial y_j^2}+\frac{1}{2}\sum_{j_1\neq j_2}\frac{\partial^2 a^2_{j_1,j_2}(y)}{\partial y_{j_1}\partial y_{j_2}}-\sum_{j=1}^{3}\frac{\partial b_j(y)}{\partial y_j}\right)f\\
&+ \sum_{j_1=1}^{3}\left(\frac{\partial a^2_{j_1,j_1}(y)}{\partial y_{j_1}}+\sum_{j_2\neq j_1}\frac{\partial a^2_{j_1,j_2}(y)}{\partial y_{j_2}}-b_{j_1}(y)\right)\frac{\partial f}{\partial y_{j_1}}+\frac{1}{2}\sum_{j_1,j_2}a^2_{j_1,j_2}(y)\frac{\partial^2 f}{\partial y_{j_1}\partial y_{j_2}}
\end{aligned}$$

Il reste donc à calculer les termes en f et $\frac{\partial f}{\partial y_j}$. Puisque :

$$b(y)=b(y_1,y_2,y_3)=\rho\begin{pmatrix}-y_1+(y_1+y_2)(y_1+y_3)\\ -y_2+(y_1+y_2)(1-y_1-y_3)\\ -y_3+(1-y_1-y_2)(y_1+y_3)\end{pmatrix}$$

nous trouvons :

$$\begin{aligned}\frac{\partial b_1}{\partial y_1} &= \rho(2y_1 + y_2 + y_3 - 1)\\ \frac{\partial b_2}{\partial y_2} &= \rho(-y_1 - y_3)\\ \frac{\partial b_3}{\partial y_3} &= \rho(-y_1 - y_2)\end{aligned}$$

et donc $\sum_{j=1}^{3} \frac{\partial b_j}{\partial y_j} = -\rho$. D'autre part rappelons que :

$$a^2(y) = a^2(y_1, y_2, y_3) = \begin{pmatrix} y_1(1-y_1) & -y_1y_2 & -y_1y_3 \\ -y_1y_2 & y_2(1-y_2) & -y_2y_3 \\ -y_1y_3 & -y_2y_3 & y_3(1-y_3) \end{pmatrix}$$

Donc on a

$$\begin{aligned}\frac{\partial a^2_{j,j}(y)}{\partial y_j} &= 1 - 2y_j\\ \frac{\partial^2 a^2_{j,j}(y)}{\partial y_j^2} &= -2\\ \frac{\partial a^2_{j_1,j_2}(y)}{\partial y_{j_2}} &= -y_{j_1}\\ \frac{\partial^2 a^2_{j_1,j_2}(y)}{\partial y_{j_1}\partial y_{j_2}} &= -1\end{aligned}$$

Le terme en f vaut par conséquent :

$$\frac{1}{2} * 3 * (-2) + \frac{1}{2} * 6 * (-1) + \rho = \rho - 6$$

et les termes en $\frac{\partial f}{\partial y_j}$ valent :

$$1 - 2y_j - y_j - y_j - b_j(y) = 1 - 4y_j - b_j(y)$$

Et on retombe finalement sur :

$$\frac{\partial f(\tau, y)}{\partial \tau} = (\rho - 6)f(\tau, y) + \sum_{j=1}^{3} \tilde{b}_j(y)\frac{\partial f(\tau, y)}{\partial y_j} + \frac{1}{2}\sum_{j_1,j_2=1}^{3} a^2_{j_1,j_2}(y)\frac{\partial^2 f(\tau, y)}{\partial y_{j_1}\partial y_{j_2}}$$

avec :

$$\tilde{b}(y) = \tilde{b}(y_1, y_2, y_3) = \begin{pmatrix} 1 - 4y_1 + \rho(y_1 - (y_1 + y_2)(y_1 + y_3)) \\ 1 - 4y_2 + \rho(y_2 + (y_1 + y_2)(1 - y_1 - y_3)) \\ 1 - 4y_3 + \rho(y_3 + (1 - y_1 - y_2)(y_1 + y_3)) \end{pmatrix}$$

B.3 Modèles de Wright-Fisher et de diffusion avec sélection

Au cours du chapitre 5, nous nous sommes principalement intéressés à des modèles neutres, c'est à dire sans sélection. Cependant nous avons dit que notre démarche pouvait s'appliquer de la même manière à des modèles avec sélection et nous avons d'ailleurs utilisé un modèle avec sélection pour l'exemple de la section 5.5. Nous montrons ici comment la sélection est intégrée au modèle de Wright-Fisher, et quelles sont les conséquences de ce changement pour la limite diffusive.

B.3.1 Modèle de Wright-Fisher à deux loci avec sélection

Le modèle que nous présentons ici est un cas particulier de celui décrit dans (Ethier et Nagylaki, 1989), ce qui nous assure la convergence vers un modèle de diffusion. Nous considérons deux loci bialléliques. Le premier locus a pour allèles Q et q, l'allèle Q ayant un avantage sélectif sur l'allèle q. Le second locus est neutre et a pour allèles 1 et 2. On note $X(t) = (X_{Q,1}(t), X_{Q,2}(t), X_{q,1}(t), X_{q,2}(t))$ le vecteur des effectifs d'haplotypes à la génération t. Comment pour le modèle neutre, nous définissons aussi le vecteur des fréquences d'haplotypes $\Pi(t) = X(t)/(2N)$. Nous supposons que la sélection s'applique aux niveaux des génotypes. Ainsi, nous introduisons pour chaque génotype $(m_1, m_2)/(n_1, n_2)$ un coefficient de viabilité $\omega_{m_1,m_2|n_1,n_2}$ qui s'écrit sous la forme

$$\omega_{m_1,m_2|n_1,n_2} = 1 - s_{m_1,m_2|n_1,n_2}, \quad m_1, n_1 = Q, q, \quad m_2, n_2 = 1, 2$$

Puisque la sélection n'agit que sur le premier locus, on prend $\forall\ m_2, n_2$

$$\begin{cases} s_{Q,m_2|Q,n_2} &= & 0 \\ s_{Q,m_2|q,n_2} &= & s_{q,m_2|Q,n_2} &= & \alpha s \\ s_{q,m_2|q,n_2} &= & s \end{cases}$$

où $s > 0$ et où $0 < \alpha < 1$ correspond au degré de dominance de l'allèle q. La loi de $X(t+1)$ sachant $X(t)$ est une loi multinomiale de paramètre $p = (p_{Q,1}, p_{Q,2}, p_{q,1}, p_{q,2})$ où

$$p_{i_1,i_2} = \frac{\sum_{m_1,m_2,n_1,n_2} \theta_{i_1,i_2:m_1,m_2|n_1,n_2} \omega_{m_1,m_2|n_1,n_2} \Pi_{m_1,m_2}(t) \Pi_{n_1,n_2}(t)}{\sum_{m_1,m_2,n_1,n_2} \omega_{m_1,m_2|n_1,n_2} \Pi_{m_1,m_2}(t) \Pi_{n_1,n_2}(t)}, \quad i_1 = Q, q, \quad i_2 = 1, 2$$

Le dénominateur correspond à la somme des coefficients de viabilité pour tous les génotypes parentaux possibles. $\theta_{i_1,i_2:m_1,m_2|n_1,n_2}$ est la probabilité de former un gamète (i_1, i_2) à partir du génotype parental $(m_1, m_2)/(n_1, n_2)$. Nous avons vu en B.1 que cette quantité vaut

$$\theta_{i_1,i_2:m_1,m_2|n_1,n_2} = \delta_{i_1,m_1}((1-c)\delta_{i_2,m_2} + c\delta_{i_2,n_2}) + \delta_{i_1,n_1}((1-c)\delta_{i_2,n_2} + c\delta_{i_2,m_2})$$

Si $s = 0$, alors $\omega_{m_1,m_2|n_1,n_2}$ ne dépend pas du génotype $(m_1, m_2)/(n_1, n_2)$ et on retrouve le modèle neutre auquel nous avons plusieurs fois fait référence dans ce manuscrit.

B.3.2 Limite diffusive

Soit $\{X^N(t)\}_{t\in\mathbb{N}}$ le processus de Wright-Fisher à deux loci défini dans la section précédente pour une population de taille N. Le vecteur $X^N(t)$ est ici le vecteur de toutes les fréquences d'haplotypes moins la dernière. D'autre part, on attribue aux haplotypes $(Q,1)$, $(Q,2)$ et $(q,1)$ les indices respectifs 1, 2 et 3. On définit, pour tout $\tau \geq 0$, $Y^N(\tau) = X^N([2N\tau])/(2N)$. On pose également $\mathbb{F}_3 = \{y \in [0,1]^J, y_1 + y_2 + y_3 \leq 1\}$.

Theorem 3 *Supposons que $\rho = \lim_{N\to+\infty} 2Nc$ et $\sigma = \lim_{N\to+\infty} 2Ns$ existent. Alors quand $N \to \infty$, le processus $\{Y^N(\tau)\}_{\tau\geq 0}$ converge en loi dans $D_{\mathbb{F}_3}[0,\infty)$ vers un processus de diffusion $\{Y(\tau)\}_{\tau\geq 0}$ à valeurs dans $\mathbb{F}_3$ et de générateur*

$$Lf(y) = \sum_{j=1}^{3} b_j(y) \frac{\partial f}{\partial y_j}(y) + \frac{1}{2} \sum_{j_1,j_2=1}^{3} a^2_{j_1,j_2}(y) \frac{\partial^2 f}{\partial y_{j_1} \partial y_{j_2}}(y)$$

avec

$$\begin{aligned} b_{Q,i_2}(y) &= \rho(-y_{Q,i_2} + y_{Q,.}y_{.,i_2}) + \sigma y_{q,.}y_{Q,i_2}(y_{Q,.} + \alpha(1 - 2y_{Q,.})), \quad i_2 = 1,2 \\ b_{q,i_2}(y) &= \rho(-y_{q,i_2} + y_{q,.}y_{.,i_2}) - \sigma y_{Q,.}y_{q,i_2}(y_{q,.} + \alpha(1 - 2y_{q,.})), \quad i_2 = 1,2 \end{aligned}$$

et

$$a^2_{i_1,i_2;i'_1,i'_2}(y) = y_{i_1,i_2}(\delta_{(i_1,i_2);(i'_1,i'_2)} - y_{i'_1,i'_2})$$

Comme dans le cas neutre, on peut appliquer l'équation projective de Kolmogorov à la densité $f(y,\tau)$ de $Y(\tau)$. En développant cette équation on obtient

$$\frac{\partial f(y,\tau)}{\partial \tau} = d(y) f(y,\tau) + \sum_{j=1}^{3} \tilde{b}_j(y) \frac{\partial f(y,\tau)}{\partial y_j} + \frac{1}{2} \sum_{j_1,j_2=1}^{3} a^2_{j_1,j_2}(y) \frac{\partial^2 f(y,\tau)}{\partial y_{j_1} \partial y_{j_2}}$$

avec : $\tilde{b}(y) =$

$$\begin{pmatrix} 1 - 4y_1 + \rho(y_1 - (y_1+y_2)(y_1+y_3)) - \sigma y_1(1-y_1-y_2)\,(y_1+y_2+\alpha(1-2y_1-2y_2)) \\ 1 - 4y_2 + \rho(y_2 + (y_1+y_2)(1-y_1-y_3)) - \sigma y_2(1-y_1-y_2)\,(y_1+y_2+\alpha(1-2y_1-2y_2)) \\ 1 - 4y_3 + \rho(y_3 + (1-y_1-y_2)(y_1+y_3)) + \sigma y_3(y_1+y_2)\,(1-y_1-y_2+\alpha(-1+2y_1+2y_2)) \end{pmatrix}$$

et

$$d(y) = \rho - 6 + \sigma(1-2\alpha)(y_1+y_2)^2 - 2\sigma(2-5\alpha)(y_1+y_2) - 2\sigma\alpha$$

Références

Ethier, S., et Nagylaki, T. (1989). Diffusion approximations of the two-locus wright-fisher model. *J. Math. Biol.*, *27*, 17-28.

Annexe C

Processus de diffusion

Nous rappelons ici les quelques définitions et propriétés utiles pour comprendre les travaux de la partie II. Pour plus de détails sur le sujet, on pourra se reporter par exemple à (Karlin et Taylor, 1981; Ethier et Kurtz, 1986; Stroock et Varadhan, 1997).

C.1 Définitions générales

On se donne dans tout ce qui suit une espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ muni d'une filtration $\{\mathcal{F}_\tau\}_{\tau \geq 0}$.

Définition 1 *Un processus stochastique de $[0, +\infty[$ dans $\mathcal{R}^d$ est dit de diffusion si et seulement si :*

1. *il possède la propriété forte de Markov*
2. *ses trajectoires sont presque sûrement continues*

Définition 2 *Un processus stochastique X possède la propriété forte de Markov relativement à $\{\mathcal{F}_\tau\}$ si et seulement si :*

1. *il est adapté à $\{\mathcal{F}_\tau\}$*
2. *$\forall T$ temps d'arrêt par rapport à $\{\mathcal{F}_\tau\}$ et $\forall h \geq 0$:*

$$\mathcal{L}(X(T+h) \mid \mathcal{F}_T) = \mathcal{L}(X(T+h) \mid X_T)$$

Proposition 3 *Si X est un processus de diffusion il vérifie notamment, $\forall x \in \mathcal{R}^d$, $\forall i, j = 1, \ldots, d$, $\forall \tau \geq 0$ et $\forall \epsilon > 0$:*

1. *Condition de Dynkin : $\frac{1}{h}\mathbb{P}(|X_i(\tau+h) - X_i(\tau)| > \epsilon \mid X(\tau) = x) \to 0$ quand $h \to 0$, uniformément sur tout compact pour x et τ.*

2. $\frac{1}{h}\mathbb{E}[X_i(\tau+h) - X_i(\tau) > \epsilon \mid X(\tau) = x] \to b_i(x,\tau)$ *quand* $h \to 0$, *uniformément sur tout compact pour* x *et* τ. $b_i(x,\tau)$ *est appelée la dérive infinitésimale du processus.*

3. $\frac{1}{h}\mathbb{E}[(X_i(\tau+h) - X_i(\tau))(X_j(\tau+h) - X_j(\tau)) > \epsilon \mid X(\tau) = x] \to a_i^2(x,\tau)$ *quand* $h \to 0$, *uniformément sur tout compact pour* x *et* τ. $a^2(x,\tau)$ *est appelée la variance infinitésimale du processus. Elle est symétrique définie positive* $\forall x \in \mathcal{R}^d$ *et* $\tau \geq 0$.

Le générateur du processus X *est l'opérateur différentiel :*

$$L_\tau f(x) = \sum_{i=1}^{3} b_i(x,\tau)\frac{\partial f(x)}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{3} a_{i,j}^2(x,\tau)\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

Le processus est dit homogène si a et b ne dépendent pas de τ.

C.2 Cas particulier

Sous certaines hypothèses sur les coefficients a et b, les processus de diffusion possèdent des propriétés supplémentaires qui en font des outils particulièrement intéressants. Nous décrivons ici ces conditions et leurs conséquences.

Soit X un processus de diffusion homogène dans $\mathcal{R}^d$ tel que $a^2(x)$ et $b(x)$ sont des fonctions C^2 à dérivées bornées sur $\mathcal{R}^d$. Supposons de plus que pour tout $x \in \mathcal{R}^d$, la matrice $a^2(x)$ soit strictement positive.

Proposition 4 *Sous les conditions énoncées ci-dessus, il existe une unique probabilité de transition* $P(\tau,x,y)$ *correspondant à une diffusion pour* a^2 *et* b. *Cette probabilité admet une densité* $f(\tau,x,y)$, *que nous appelons densité de transition, telle que pour tout* $\tau \geq 0$ *:*

$$P(\tau,x,A) = \int_{y\in A} f(\tau,x,y)dy$$

$f(\tau,x,y)$ *est* C^0 *et strictement positive sur* $[0,+\infty[\times\mathcal{R}^d \times \mathcal{R}^d$, C^2 *en* x *et* C^1 *en* τ. *Elle vérifie, en tant que fonction de* τ *et* x, *l'équation rétrograde de Kolmogorov :*

$$\frac{\partial f(\tau,x,y)}{\partial \tau} = Lf(\tau,x,y) \qquad \tau > 0$$

et en tant que fonction de τ *et* y, *l'équation rétrograde de Kolmogorov :*

$$\frac{\partial f(\tau,x,y)}{\partial \tau} = L^* f(\tau,x,y) \qquad \tau > 0$$

où L^* est l'opérateur adjoint de L défini par :

$$L^* f(y) = -\sum_{i=1}^{3} \frac{\partial(b_i(y)f(y))}{\partial y_i} + \frac{1}{2}\sum_{i,j=1}^{3} \frac{\partial^2(a_{i,j}^2(y)f(y))}{\partial y_i \partial y_j}$$

Remarque 6 *Une conséquence directe de l'équation rétrograde de Kolmogorov est que pour toute fonction $g : \mathcal{R}^d \to \mathcal{R}^d$ bornée et continue par morceaux, la fonction $u(\tau, x) = \mathbb{E}_x[g(X)]]$ vérifie l'équation :*

$$\frac{\partial u(\tau, x)}{\partial \tau} = Lu(\tau, x) \ \forall x \in \mathcal{R}^d, \forall \tau > 0$$

Références

Ethier, S., et Kurtz, T. (1986). *Markov processes. characterization and convergence.* John Wiley and Sons, Inc.

Karlin, S., et Taylor, H. (1981). *A second course in stochastic processes.* Academic Press, Inc.

Stroock, D., et Varadhan, S. (1997). *Multidimensionnal diffusion processes.* Springer-Verlag.