VALÉRIE JOMPHE

COMPARAISON DE LA PUISSANCE DE TESTS DE DÉSÉQUILIBRE DE LIAISON DANS LES ÉTUDES GÉNÉTIQUES

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en statistique
pour l'obtention du grade de Maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL
QUÉBEC

2006

Résumé

L'identification du gène responsable d'une maladie peut être facilitée par des méthodes statistiques telles que des études d'association basées sur le déséquilibre de liaison. Différentes stratégies d'analyse sont possibles pour ce type d'étude. Comme pour les tests d'association classiques, un devis d'échantillonnage de cas-témoins peut être utilisé. Un deuxième devis possible est l'échantillonnage de trios. On peut également choisir d'étudier l'association allélique ou haplotypique des marqueurs génétiques sélectionnés. La présente étude vise à comparer par voie de simulation la puissance de tests de déséquilibre de liaison selon la stratégie d'analyse choisie. Dans un premier temps, on s'est intéressé à la comparaison des devis d'échantillonnage cas-témoins et trios ; dans un deuxième temps, on a comparé les approches allélique et haplotypique.

Avant-propos

Ce mémoire de maîtrise est le résultat de plusieurs mois de travail durant lesquels j'ai vécu des moments de bonheur où régnait un sentiment de satisfaction et de dépassement de soi, mais aussi des moments de stress et de découragement. Heureusement, j'ai pu compter sur l'aide et le soutien de plusieurs personnes qui ont contribué de près ou de loin à la réalisation de ce projet.

J'aimerais tout d'abord remercier ma co-directrice Mme Chantal Mérette, directrice du Laboratoire de biostatistique et de psychiatrie génétique du Centre de recherche Université Laval Robert-Giffard (CRULRG), qui a proposé cet intéressant projet lié à un domaine de recherche passionnant, la statistique génétique. Elle m'a guidée tout au long de ce projet tout en me faisant confiance et en me laissant une certaine liberté ; je dois dire que j'ai appris beaucoup. De plus, elle m'a accordé une place dans son équipe de recherche au CRULRG où j'ai pu prendre de l'expérience de travail en tant que statisticienne tout en travaillant sur mon projet de maîtrise.

Je suis également reconnaissante envers mon directeur M. Christian Genest, professeur au Département de mathématiques et de statistique de l'Université Laval, qui a cru en moi et m'a encouragée à poursuivre des études de deuxième cycle. Je lui suis redevable de nombreux conseils, à la fois judicieux et très constructifs, pour la rédaction de mon mémoire et la présentation des parties théoriques qu'il comporte.

Je désire également exprimer ma gratitude envers Mme Aurélie Labbe, professeure au Département de mathématiques et de statistique de l'Université Laval, dont le cours de statistique génétique m'aura permis de consolider mes connaissances dans un nouveau domaine d'étude en plus de les approfondir. D'ailleurs, deux des exemples présentés dans ce mémoire sont tirés de son cours. Je tiens aussi à la remercier de m'avoir rassurée et encouragée en me faisant part de sa propre expérience.

Je dois souligner par ailleurs la contribution de mes collègues de travail. Je pense entre autres à Nathalie Savard et Julie Béliveau qui, ayant terminé leur mémoire

quelques mois avant moi, ont su me conseiller, que ce soit au niveau de la rédaction ou de la programmation de mes simulations. Je n'oublie pas non plus Annie Labbé, qui m'a encouragée dans les dernières, et non les moindres, semaines de rédaction et qui a répondu à mes questions en matière de français.

Le Conseil de recherches en sciences naturelles et en génie du Canada a financé une partie de mes travaux de recherche en m'octroyant une bourse d'études supérieures. Mes directeurs de recherche ont également contribué au financement de ce projet en m'accordant une partie de leurs fonds de recherche. Je leur en sais gré.

En terminant, je tiens à remercier toute ma famille qui m'encourage depuis toujours. Je pense plus particulièrement à mes parents, Jean et Diane, qui malgré les mille kilomètres de route qui nous séparent ont toujours été là pour moi, autant pour me soutenir financièrement que moralement, et ce, tout au long de mes études. Chers parents, j'apprécie tout ce que vous avez fait pour moi. Je sais que vous êtes fiers de ce que j'ai réalisé et je tiens à ce que vous sachiez qu'une grande partie du mérite vous revient. Sans vous, je ne serais pas celle que je suis aujourd'hui.

Derrière chaque femme se cache un homme qui sait écouter, réconforter et faire preuve de patience. Cet homme est pour moi mon amoureux Jason : je te remercie pour avoir enduré mes hauts et mes bas... je sais que ça n'a pas toujours été facile ! Tu as toujours su m'encourager et me faire voir le bon côté des choses. Je te remercie de m'avoir permis de garder un équilibre dans ma vie. Ta présence m'aura été indispensable.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

En ce début du 21^{e} siècle, la recherche en génétique humaine est en plein essor. Les gènes responsables de certaines pathologies ont déjà été identifiés. D'autres maladies génétiques plus complexes représentent toutefois un défi de taille pour les chercheurs. De nouvelles méthodes d'identification de ces gènes sont continuellement proposées. La statistique apporte une contribution importante à ce domaine de recherche.

Une des méthodes d'analyse statistique fréquemment utilisées est l'étude d'association basée sur le déséquilibre de liaison (DL) génétique. Plusieurs stratégies d'analyse sont disponibles pour ce type d'étude, en fonction du devis d'échantillonnage et de l'approche analytique envisagée. Deux devis d'échantillonnage sont couramment utilisés : le premier, appelé devis de cas-témoins, est classique dans les études d'association ; le second, qui est spécifique aux travaux en génétique, est fondé sur l'échantillonnage de « trios », à savoir des familles constituées d'un enfant et de ses deux parents. Par ailleurs, l'analyse statistique peut porter sur l'association allélique ou sur l'association haplotypique des marqueurs génétiques considérés.

L'étude de l'efficacité de ces différentes stratégies d'analyse revêt une grande importance, autant du point de vue statistique qu'opérationnel. Il est souhaitable d'identifier les méthodes les plus efficaces et les contextes dans lesquels certaines approches s'avèrent préférables. Les chercheurs intéressés par les études d'association génétique disposeraient alors d'un outil pratique de planification et de réalisation de leurs travaux.

L'objectif de ce mémoire est de contribuer à l'avancement des connaissances concernant l'efficacité des diverses méthodes statistiques d'analyse dans le cadre d'études d'association. On cherchera plus spécifiquement à comparer la puissance statistique des différentes stratégies d'analyse par voie de simulation.

En premier lieu, on comparera la puissance des deux devis d'échantillonnage, cas-témoins et trios. Parce qu'il présente l'avantage de fournir une information génétique supplémentaire, c'est-à-dire les génotypes des parents, l'échantillonnage de trios facilite l'identification des haplotypes et devrait en principe être préférable à l'échantillonnage de cas-témoins. C'est d'autant plus plausible que ce mode d'échantillonnage assure un appariement parfait des allèles cas et des allèles témoins. Par opposition, l'échantillonnage de cas-témoins demande des précautions supplémentaires lors de la sélection des témoins, lesquels doivent être similaires aux cas pour éviter les problèmes de stratification de la population.

En deuxième lieu, on cherchera à comparer la puissance des deux approches analytiques, allélique et haplotypique. On s'attend a priori à ce que l'approche haplotypique soit avantageuse, puisqu'elle permet de combiner l'information de plusieurs marqueurs à la fois, de tenir compte de la structure de DL qui existe entre les marqueurs étudiés et de réduire le nombre de tests.

Le chapitre 2 introduit les concepts de base en génétique nécessaires à la compréhension du mémoire. Au chapitre 3, le principe général des études d'association appliquées à la génétique est présenté. C'est dans ce chapitre que le déséquilibre de liaison est défini ; les différentes stratégies d'analyse possibles y sont également exposées.

Des tests d'association pour les échantillons de cas-témoins sont détaillés au chapitre 4. Le chapitre 5 présente les résultats d'études d'association réalisées dans des échantillons de cas-témoins dans le cadre de la recherche des gènes de vulnérabilité à la schizophrénie. Des tests d'association pour les échantillons de trios sont décrits au chapitre 6. Au chapitre 7, les avantages et inconvénients des différentes stratégies d'analyse sont énumérés. On y présente également des études récentes concernant la comparaison de ces méthodes.

La structure de l'étude de Monte-Carlo réalisée dans le cadre de ce mémoire est présentée au chapitre 8. Les résultats obtenus et les principales conclusions qui en découlent sont consignés au chapitre 9. Quelques remarques regroupées au chapitre 10 complètent le tout.

Chapitre 2

Introduction à la génétique

On appelle génétique la science qui étudie la transmission des caractères d'une génération à l'autre. Un des principaux objectifs de la recherche en génétique humaine est l'identification des gènes responsables des maladies. La connaissance de ces gènes aide au développement de nouvelles méthodes de prévention, de diagnostic et de traitement. La recherche en génétique est possible grâce à la collaboration de scientifiques provenant de plusieurs disciplines, notamment la biologie moléculaire, la médecine, l'informatique et la statistique.

La génétique connaît du succès. Les gènes de certaines maladies ont déjà été localisés. C'est ainsi par exemple qu'en 1993, on a identifié un gène de vulnérabilité à la maladie de Huntington (The Huntington's Disease Collaborative Research Group, 1993). BRCA1 et BRCA2 sont reconnus comme étant des gènes qui prédisposent au développement de certaines formes de cancer du sein (Miki et coll., 1994; Wooster et coll., 1995). Les pathologies complexes, telles que la schizophrénie, la maladie bipolaire et le diabète, présentent toutefois un défi de taille pour la génétique. Une hétérogénéité pouvant impliquer une interaction entre des facteurs génétiques et environnementaux serait à l'origine de celles-ci. L'effet de leurs gènes serait donc plus modéré et plus difficile à détecter.

2.1 Notions de base en génétique

Le *génome* représente l'ensemble du matériel génétique, constitué de molécules d'acide désoxyribonucléique (ADN), transmis de génération en génération. Chez l'hu-

main, l'information génétique est assemblée en vingt-trois paires de chromosomes que l'on retrouve dans le noyau de chacune de nos cellules. Voir les figures 2.1 et 2.2.

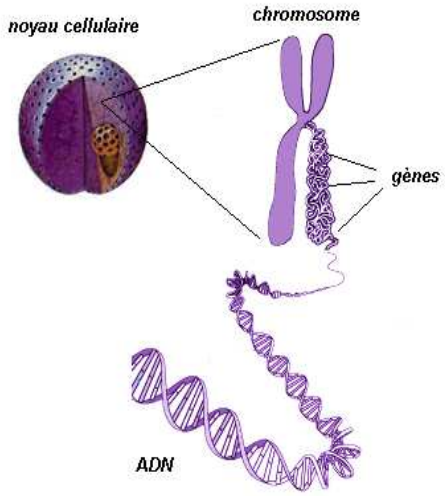


FIG. 2.1 – *Chromosome, gène et ADN (Lampron, 2006)*

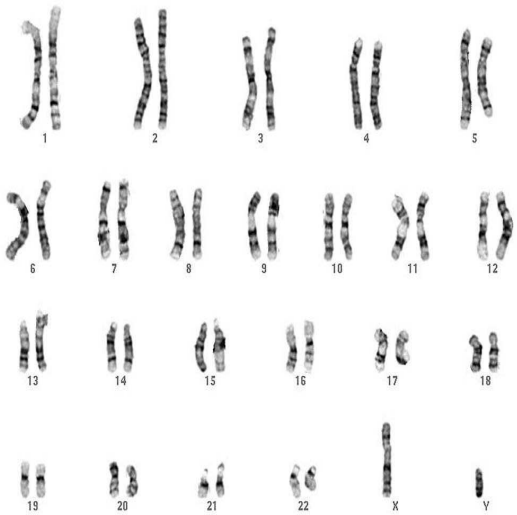


FIG. 2.2 – *Vingt-trois paires de chromosomes chez l'humain (Chromosome 3q Registry, 2006)*

Le terme *chromosome* fait référence à la forme condensée de l'ADN. Une séquence d'environ trois milliards de paires de bases constitue l'ensemble de ces chromosomes. Les *bases* sont les principales molécules qui composent l'ADN. Il en existe quatre : l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Chaque base est combinée à un sucre et à un groupement phosphate, ce qui représente un *nucléotide*. Les *gènes*, qui sont des fragments d'ADN, déterminent les caractéristiques de chaque individu : l'apparence physique, la vulnérabilité à certaines maladies, la réponse à l'exposition aux facteurs environnementaux, etc.

Depuis 2003, on connaît la séquence complète du génome humain. On estime à environ 20 000 à 25 000 le nombre de gènes (Human Genome Project, 2006). Les chercheurs des quatre coins du monde disposent maintenant d'une source d'information très riche. L'ère de la génétique n'en est qu'à ses débuts.

On appelle *locus* une position particulière sur un chromosome. Les différentes formes existantes à un locus donné sont appelées *allèles*. Considérons par exemple le gène XYZ au locus D1S02 du chromosome 1 de la figure 2.3. Ce gène XYZ possède trois allèles : A, B et C.

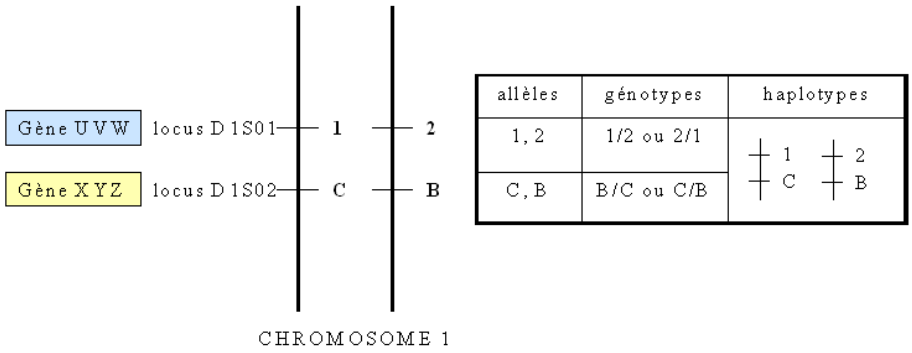


FIG. 2.3 – *Un exemple d'information génétique aux loci D1S01 et D1S02*

Le *génotype* d'un individu est la combinaison des allèles de la paire de chromosomes à un ou plusieurs loci donnés. Un allèle vient du père et l'autre de la mère. Généralement, on note le génotype en séparant les allèles des deux chromosomes par une barre oblique (/). Si les deux allèles sont identiques, par exemple A/A, on dit que le génotype est *homozygote*. Si les allèles sont différents, par exemple A/B, le génotype est dit *hétérozygote*. Supposons que le nombre d'allèles est k. Le nombre de génotypes possible s'obtient en additionnant le nombre de génotypes hétérozygotes, $\binom{k}{2} = k(k-1)/2$, et le nombre de génotypes homozygotes, $\binom{k}{1} = k$.

Pour le gène XYZ, six génotypes sont possibles :

$$A/A,\ A/B,\ A/C,\ B/B,\ B/C \text{ et } C/C.$$

L'ensemble des allèles de deux loci ou plus situés sur un même chromosome est appelé un *haplotype*. Au locus D1S01 représenté à la figure 2.3, on retrouve le gène UVW, qui se situe juste au-dessus du gène XYZ. Deux allèles sont possibles pour ce gène, les allèles 1 et 2. Le nombre d'haplotypes possibles est égal au produit du nombre d'allèles pour chaque gène, soit $2 \times 3 = 6$. Les six haplotypes en question sont : $1A, 1B, 1C, 2A, 2B, 2C$.

2.2 Diversité génétique

Quand on cherche les gènes responsables d'une maladie, on s'intéresse aux différences dans le code génétique des individus. Quelle séquence particulière prédispose à dévelop-

per une maladie ? Ces variations sont appelées *polymorphismes*. Elles sont le résultat de mutations. Une *mutation* correspond à une modification soudaine et transmissible de la séquence d'ADN. Par exemple, le changement d'une base par une autre ou encore l'ajout ou la suppression d'une base. Des mutations auraient permis l'évolution de l'espèce humaine. Des mutations seraient également responsables de l'apparition de certaines maladies.

Une autre source de diversité génétique entre individus est la recombinaison. La *recombinaison* est un phénomène résultant du mélange de matériel génétique qui se produit par enjambement entre chromosomes ; voir figure 2.4. Elle survient au cours de la *méïose*, le processus de formation des spermatozoïdes et des ovules. Chaque chromosome a alors la possibilité d'échanger une partie d'ADN avec son chromosome homologue. Plus deux loci sont proches l'un de l'autre, moins il y a de chances qu'il y ait une recombinaison entre les deux.

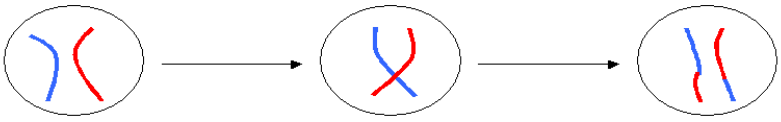


FIG. 2.4 – *Un exemple de recombinaison à la suite d'un enjambement entre des chromosomes*

En génétique, on étudie des *marqueurs* qui sont des polymorphismes dont la fonction n'est pas nécessairement connue, mais dont on connaît la position exacte sur le génome. Un type de marqueurs couramment utilisé est le polymorphisme de nucléotide simple (SNP) qui résulte de la modification d'un seul nucléotide tel qu'illustré à la figure 2.5.

Pour qu'un SNP soit utile en génétique, la fréquence de l'allèle résultant de la modification doit être supérieure à 1% dans la population. On estime à 10 millions le nombre de SNP chez l'humain (International HapMap Project, 2006). Le projet HapMap (The International HapMap Consortium, 2005) a permis de cartographier la diversité génétique de l'espèce humaine. Quatre populations originaires de diverses parties du monde ont été étudiées : la Chine, les États-Unis, le Japon et le Nigéria. L'objectif de ce projet : rendre accessible aux chercheurs de ce monde l'information sur les variations génétiques les plus fréquentes chez l'humain. HapMap constitue un outil précieux pour la recherche en génétique.

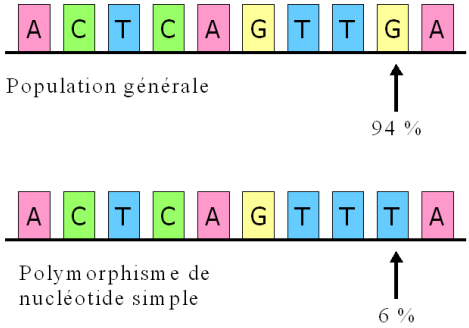


FIG. 2.5 – *Un exemple de SNP (Genome News Network, 2006)*

2.3 L'équilibre de Hardy–Weinberg

L'équilibre de Hardy–Weinberg est un principe fondamental en génétique des populations qui soutient que les fréquences génotypiques à un locus donné demeurent constantes de génération en génération si les cinq conditions suivantes sont respectées :

1. la population est de grande taille ;
2. il n'y a pas de sélection naturelle[1] ;
3. il n'y a pas de mutation ;
4. il n'y a pas de migration, c'est-à-dire qu'aucune copie allélique n'est apportée de l'extérieur ;
5. les unions sont aléatoires, le choix d'un partenaire ne dépend pas de son génotype.

Soit un locus à deux allèles, disons A et a, avec fréquences alléliques respectives p et q. Trois génotypes peuvent être observés : A/A, A/a et a/a. On s'attend aux fréquences génotypiques suivantes si la population est en équilibre de Hardy–Weinberg :

$$f\,(A/A) = p^2, \quad f\,(A/a) = 2pq, \quad f\,(a/a) = q^2.$$

[1] Concept de la théorie de l'évolution de Darwin qui stipule que les variations entre individus les plus propices à la survie et à la reproduction se transmettront mieux que les autres (Encyclopédie Wikipédia en ligne).

2.3.1 Vérification du postulat d'équilibre de Hardy–Weinberg

Supposons que l'on a génotypé un échantillon de n individus au locus présenté à la section précédente. Le nombre, x, d'allèles vaut 2. Le nombre, m, de génotypes vaut 3. Le nombre total de génotypes dans l'échantillon est égal au nombre, n, d'individus alors que le nombre total d'allèles est égal à deux fois le nombre d'individus, soit $2n$.

Les fréquences génotypiques relatives sont : $n_{A/A}$, $n_{A/a}$ et $n_{a/a}$. La somme des trois est égale au nombre n d'individus. Les fréquences alléliques relatives, dont la somme est égale à $2n$, sont données par :

$$n_A = 2n_{A/A} + n_{A/a}, \quad n_a = 2n_{a/a} + n_{A/a}.$$

Pour vérifier l'hypothèse du respect de l'équilibre de Hardy–Weinberg, on effectue simplement un test du khi-deux. À cette fin, on calcule la statistique

$$X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i},$$

où O_i et E_i représentent respectivement les fréquences observée et espérée de chaque génotype, telles que précisées dans le tableau 2.1. Mentionnons que les fréquences espérées sont calculées sous l'hypothèse H_0 de respect de l'équilibre de Hardy–Weinberg.

	génotype		
	A/A	A/a	a/a
O_i	$n_{A/A}$	$n_{A/a}$	$n_{a/a}$
E_i	np^2	$2npq$	nq^2

TAB. 2.1 – Fréquences observées et espérées des trois génotypes

Sous l'hypothèse nulle d'équilibre de Hardy–Weinberg, X^2 obéit asymptotiquement à une loi du khi-deux à k degrés de liberté, où

$$k = \begin{cases} m-1, & \text{si on connaît les fréquences d'allèles ;} \\ m-x, & \text{si on doit estimer les fréquences des } x \text{ allèles.} \end{cases}$$

La procédure `allele` du module `genetics` du logiciel SAS et la fonction `hwe.chisq` du logiciel R permettent de vérifier si l'équilibre de Hardy–Weinberg est respecté.

2.3.2 Test exact

Lorsque certaines des fréquences espérées sont inférieures à 5, il est préférable d'utiliser un test exact tel que proposé par Guo et Thompson (1992).

Sous H_0, le vecteur $\left(n_{A/A}, n_{A/a}, n_{a/a}\right)$ des fréquences génotypiques absolues a une distribution multinomiale avec probabilités respectives $p^2, 2pq$ et q^2 :

$$P\left(n_{A/A}, n_{A/a}, n_{a/a}\right) = \frac{n!}{n_{A/A}! n_{A/a}! n_{a/a}!} \left(p^2\right)^{n_{A/A}} (2pq)^{n_{A/a}} \left(q^2\right)^{n_{a/a}}.$$

De même, sous H_0, le vecteur (n_A, n_a) des fréquences alléliques absolues a une distribution multinomiale avec probabilités respectives p et q :

$$P\left(n_A, n_a\right) = \frac{2n!}{n_A! n_a!} (p)^{n_A} (q)^{n_a}.$$

La statistique du test correspond à la probabilité d'observer les fréquences génotypiques $\left(n_{A/A}, n_{A/a}, n_{a/a}\right)$ sachant que les fréquences alléliques sont (n_A, n_a). Elle s'obtient comme suit :

$$\begin{aligned} P\left\{\left(n_{A/A}, n_{A/a}, n_{a/a}\right) \mid (n_A, n_a)\right\} &= \frac{P\left\{\left(n_{A/A}, n_{A/a}, n_{a/a}\right) \bigcap (n_A, n_a)\right\}}{P\left(n_A, n_a\right)} \\ &= \frac{P\left(n_{A/A}, n_{A/a}, n_{a/a}\right)}{P\left(n_A, n_a\right)} \\ &= \frac{n!}{(2n)!} \frac{n_A! n_a!}{n_{A/A}! n_{A/a}! n_{a/a}!} 2^{n_{A/a}}. \end{aligned}$$

Le seuil observé du test se calcule par un processus de permutation des données de l'échantillon. Les $2n$ allèles sont permutés aléatoirement pour former tous les ensembles possibles de n génotypes. Le seuil observé correspond à la proportion des statistiques calculées sur les données permutées dont la valeur est supérieure à celle de la statistique calculée sur les données originelles.

La version exacte du test de Hardy–Weinberg peut être effectuée au moyen de la procédure `allele` de SAS ou de la fonction `hwe.exact` de R.

Étant donné que la plupart des tests utilisés en génétique supposent que la population étudiée est en équilibre de Hardy–Weinberg, il est préférable de faire la vérification

avant toute analyse. Sinon, l'interprétation des résultats risque d'être faussée. Il est important de mentionner qu'on teste généralement l'équilibre de Hardy–Weinberg chez les cas et les témoins séparément. On exige que l'équilibre soit respecté au moins chez les témoins car, si le marqueur étudié est associé à la maladie, l'équilibre chez les cas pourrait être affecté par cette association.

2.4 Localiser les gènes responsables des maladies

Il existe deux principaux types d'études en génétique pour identifier le ou les gènes responsables d'une maladie :

- l'étude de liaison ;
- l'étude d'association.

Les études de liaison exploitent l'information génétique contenue dans les familles. On dit qu'il y a liaison entre deux loci si leurs allèles sont transmis ensemble dans plus de 50% des cas. En fait, deux loci sont liés s'ils sont situés proches l'un de l'autre, ce qui diminue les chances de recombinaison et, par le fait même, augmente les chances que leurs allèles soient transmis ensemble. Comme on ne connaît pas le locus de la maladie, on cherche des marqueurs situés dans son voisinage. Si l'allèle d'un marqueur est fréquemment transmis aux personnes malades dans les familles, on dira que ce marqueur est lié à la maladie. La recherche du gène responsable pourra alors se concentrer autour de ce marqueur.

Les études d'association ont pour but d'identifier un marqueur pour lequel un allèle est plus fréquent chez les sujets malades que chez les sujets sains. On dit alors que cet allèle est associé à la maladie. Il est possible d'effectuer ces études à partir d'échantillons de cas-témoins, mais également à partir d'échantillons de familles.

Il est important de faire la distinction entre les deux types d'études. L'association est une hypothèse plus forte que celle de la liaison. L'allèle associé au trait est le même pour toute la population. En liaison, différents allèles peuvent être liés à la maladie dans différentes familles. S'il y a association, il y a nécessairement liaison. Cependant, la liaison n'entraîne pas l'association. Les figures 2.6 et 2.7, gracieuseté de la professeure Aurélie Labbe, illustrent bien cette distinction. Dans ces figures comme dans toutes les suivantes, un carré représente un homme et un cercle une femme. De plus, les sujets malades sont représentés par des figures pleines.

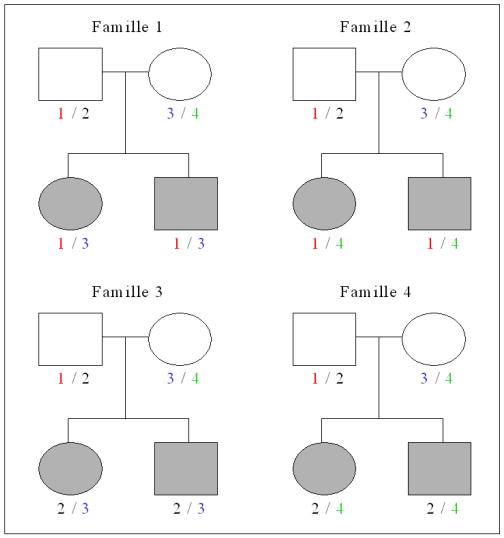


FIG. 2.6 – *Liaison sans association*

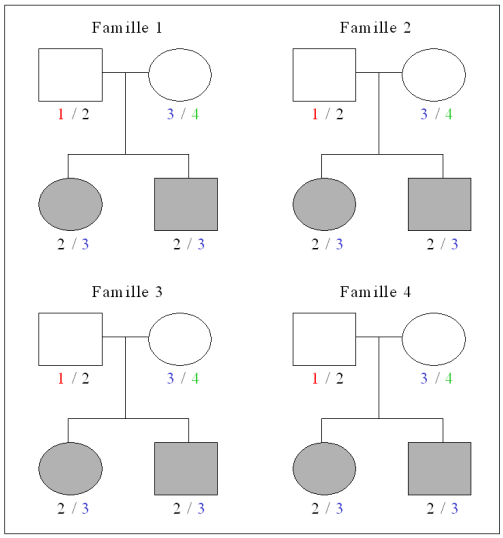


FIG. 2.7 – *Liaison avec association*

La figure 2.6 présente des familles pour lesquelles des allèles différents sont liés à la maladie (familles 1 et 2 : allèle 1, familles 3 et 4, allèle 2, familles 1 et 3 : allèle 3, familles 2 et 4 : allèle 4). On peut dire qu'il y a présence de liaison, mais absence d'association. Dans la figure 2.7, les allèles 2 et 3 sont liés à la maladie, et ce pour les quatre familles. Il y a donc présence de liaison et d'association.

Généralement, on utilise les études de liaison pour identifier des zones du génome candidates pour le gène responsable de la maladie, puis les études d'association pour préciser l'emplacement du gène (« fine-mapping »). Risch et Merikangas (1996) signalent que les études de liaison sont surtout efficaces pour identifier des gènes qui ont un effet important. De plus, ils ont démontré que les études d'association sont plus puissantes que les études de liaison quand l'effet du gène est plus modeste, comme c'est le cas avec des maladies complexes.

Le présent travail s'attardera en particulier à l'étude d'association qui sera présentée au prochain chapitre.

Chapitre 3

Survol des études d'association génétique

Les études d'association ont permis d'identifier des gènes de vulnérabilité à certaines maladies complexes. Mentionnons entre autres le gène APOE qui prédispose à la maladie d'Alzheimer (Corder et coll., 1993), le gène NOD2/CARD15 qui contribue au développement de la maladie de Crohn (Hugot et coll., 2001) et le gène KCNJ11 qui est associé au diabète de type 2 (Gloyn et coll., 2003).

Le principe des études d'association génétique se résume à identifier un marqueur pour lequel un allèle est significativement plus fréquent chez les gens malades que chez les bien portants.

Il est important de noter toutefois qu'une association entre un marqueur et une maladie peut se présenter de trois façons :

- **association directe** :
  le marqueur est directement lié à la maladie, c'est-à-dire qu'il cause la maladie ou augmente les risques de la développer ;
- **association indirecte** :
  le marqueur ne cause pas directement la maladie, mais est situé proche du locus de la maladie ;
- **fausse association** :
  l'association peut être confondue avec une stratification dans la population (détails à la section 4.3).

3.1 Le déséquilibre de liaison

3.1.1 Définition du déséquilibre de liaison

Les études d'association sont fondées sur le principe de déséquilibre de liaison (DL). Le DL réfère à l'association non aléatoire entre les allèles de deux ou plusieurs loci. Plus deux loci sont situés proches l'un de l'autre, moins leurs allèles ont de chances d'être transmis de façon indépendante. On dit alors qu'ils sont en DL. Des loci très près l'un de l'autre seront en plus fort DL que des loci plus éloignés. Ce concept est très utile pour la recherche en génétique. En effet, une association entre un marqueur et une maladie suggère la présence d'un DL entre les deux. On présume alors que le marqueur est à proximité du locus de la maladie, ce qui peut faciliter l'identification du gène.

Rappelons que des mutations sont souvent à l'origine de l'apparition des maladies. Quand une nouvelle mutation survient sur un chromosome, elle sera transmise aux générations suivantes avec tous les autres marqueurs de ce même chromosome, à moins qu'il n'y ait recombinaison. La mutation est en DL complet avec les loci adjacents tant qu'il ne se produit pas de recombinaison.

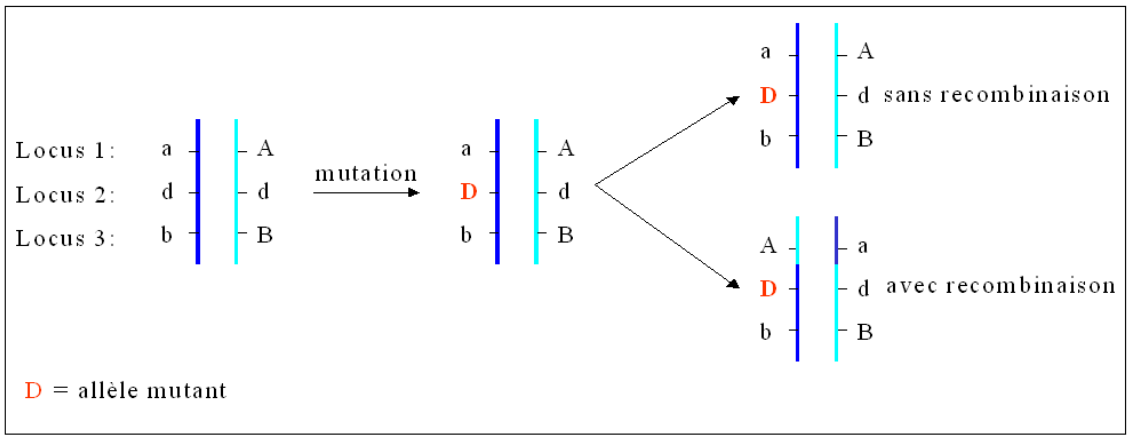


FIG. 3.1 – *Un exemple de mutation avec et sans recombinaison*

Dans l'exemple de la figure 3.1, les allèles a et b sont transmis avec l'allèle mutant D s'il n'y a pas de recombinaison. S'il en survient une, des allèles différents seront hérités avec D selon l'emplacement de la recombinaison. Dans l'exemple, une recombinaison entre les loci 1 et 2 se produit. L'allèle A est alors transmis avec l'allèle mutant plutôt que l'allèle a.

Avec le temps, le DL diminue en raison de ces recombinaisons. Or, plus la distance entre deux marqueurs est petite, moins il y a de chances qu'il y ait recombinaison entre les deux et le DL diminue moins rapidement. On appelle *taux de recombinaison*, θ, la probabilité qu'il y ait recombinaison entre deux loci. Le DL dépend donc du temps t, en nombre de générations depuis l'apparition de la mutation, et du taux θ de recombinaison :

$$\mathcal{D}_t = \mathcal{D}_0(1-\theta)^t.$$

Ici, $\mathcal{D}_0$ représente le DL au moment de la mutation. La variation de $\mathcal{D}_t$ en fonction du temps t est illustrée à la figure 3.2 pour différents choix de taux de recombinaison θ.

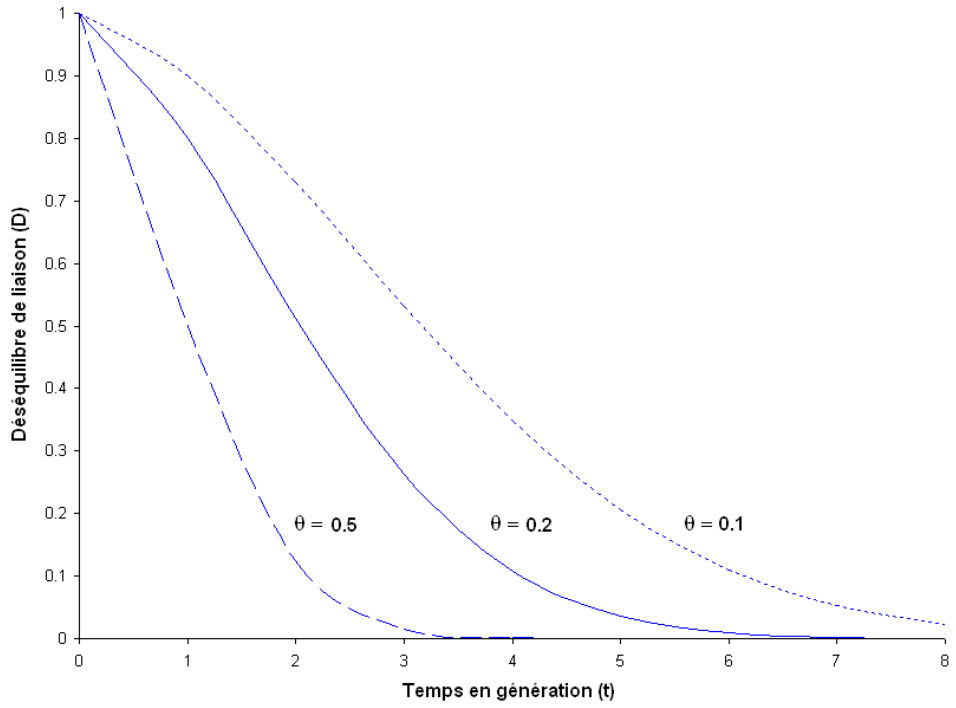


FIG. 3.2 – *Le déséquilibre de liaison en fonction du temps et du taux de recombinaison*

3.1.2 Les mesures de déséquilibre de liaison

On mesure la force du déséquilibre de liaison entre deux marqueurs à l'aide du coefficient de déséquilibre de liaison, $\mathcal{D}$. Soient deux marqueurs, $M1$ et $M2$, possédant respectivement les allèles A et a et les allèles B et b. Le coefficient de déséquilibre de liaison entre $M1$ et $M2$ correspond à la différence entre la proportion d'haplotypes AB (ou ab) observée et celle espérée sous l'hypothèse d'indépendance. Si les deux marqueurs

sont indépendants, on s'attend à ce que la proportion d'haplotypes AB soit égale au produit des fréquences d'allèles A et B, c'est-à-dire

$$\mathcal{D} = \mathrm{P}(AB) - \mathrm{P}(A)\mathrm{P}(B) = \mathrm{P}(ab) - \mathrm{P}(a)\mathrm{P}(b).$$

Ainsi, plus $\mathcal{D}$ est élevé, plus les marqueurs sont en déséquilibre de liaison.

Des standardisations de $\mathcal{D}$ ont été proposées afin d'avoir des coefficients compris entre -1 et 1. Les mieux connues sont le coefficient de corrélation r et le $\mathcal{D}'$ de Lewontin (1964). Ces indices se calculent respectivement comme suit :

$$r = \frac{\mathcal{D}}{\sqrt{\mathrm{P}(A)\mathrm{P}(a)\mathrm{P}(B)\mathrm{P}(b)}}\,, \quad \mathcal{D}' = \frac{\mathcal{D}}{\mathcal{D}_{\max}}\,,$$

où

$$\mathcal{D}_{\max} = \begin{cases} \min\left\{\mathrm{P}(A)\mathrm{P}(b), \mathrm{P}(a)\mathrm{P}(B)\right\}, & \text{si } \mathcal{D} > 0\,; \\ \min\left\{\mathrm{P}(A)\mathrm{P}(B), \mathrm{P}(a)\mathrm{P}(b)\right\}, & \text{si } \mathcal{D} < 0. \end{cases}$$

Comment obtient-on la formule pour la valeur maximale de $\mathcal{D}$? Les détails de ce calcul, présentés dans le livre de Weir (1990), sont les suivants. La fréquence d'un haplotype ne peut être ni négative ni supérieure aux fréquences d'allèles qui le composent. En d'autres termes, on doit avoir

$$0 \leq \mathrm{P}(AB) \leq \min\{\mathrm{P}(A), \mathrm{P}(B)\}.$$

L'ensemble des valeurs possibles du coefficient de déséquilibre de liaison est donc restreint. Si $\mathcal{D}$ est positif, alors le fait que

$$-\mathrm{P}(A)\mathrm{P}(B) \leq \mathcal{D} \leq \mathrm{P}(A) - \mathrm{P}(A)\mathrm{P}(B)$$

et que

$$-\mathrm{P}(A)\mathrm{P}(B) \leq \mathcal{D} \leq \mathrm{P}(B) - \mathrm{P}(A)\mathrm{P}(B)$$

entraîne successivement

$$\mathcal{D} \leq \mathrm{P}(A)\{1 - \mathrm{P}(B)\} \quad \text{et} \quad \mathcal{D} \leq \mathrm{P}(B)\ \{1 - \mathrm{P}(A)\},$$

d'où

$$\mathcal{D} \leq \mathrm{P}(A)\mathrm{P}(b) \quad \text{et} \quad \mathcal{D} \leq \mathrm{P}(B)\mathrm{P}(a),$$

ce qui permet de conclure que

$$\mathcal{D}_{\max} = \min\left\{\mathrm{P}(A)\mathrm{P}(b), \mathrm{P}(a)\mathrm{P}(B)\right\}.$$

Une démarche semblable conduit au résultat pour un $\mathcal{D}$ négatif.

Mentionnons que la procédure `allele` de SAS et la fonction `ld` du logiciel R permettent d'obtenir les différentes mesures de DL entre les marqueurs étudiés.

3.1.3 Exemples de coefficients de déséquilibre de liaison

Les exemples numériques qui suivent illustrent le calcul des coefficients de déséquilibre de liaison $\mathcal{D}$, $\mathcal{D}'$ et r dans différentes situations. Les deux premiers cas correspondent à des DL fort et faible. Les deux suivants permettent de distinguer entre les cas où $|\mathcal{D}'| = 1$, selon que $|r|$ est lui-même égal à 1 ou inférieur à 1 ; on parle alors respectivement de situations de DL total et complet.

1) DL fort

		$M2$		
		B	b	
$M1$	A	0.8172	0.0737	0.8909
	a	0.0070	0.1022	0.1092
		0.8242	0.1759	1

$$\begin{aligned}
\mathcal{D}(AB) &= 0.8172 - 0.8909 \times 0.8242 = 0.0829, \\
\mathcal{D}'(AB) &= \frac{0.0829}{\min(0.8909 \times 0.1759, 0.1092 \times 0.8242)} = 0.9211, \\
r(AB) &= \frac{0.0829}{\sqrt{0.8909 \times 0.1092 \times 0.8242 \times 0.1759}} = 0.6980.
\end{aligned}$$

2) DL faible

		$M2$		
		B	b	
$M1$	A	0.81	0.0892	0.8992
	a	0.0851	0.0157	0.1008
		0.8951	0.1049	1

$$\begin{aligned}
\mathcal{D}(AB) &= 0.81 - 0.8992 \times 0.8951 = 0.0051, \\
\mathcal{D}'(AB) &= \frac{0.0051}{\min(0.8992 \times 0.1049, 0.1008 \times 0.8951)} = 0.0565, \\
r(AB) &= \frac{0.0051}{\sqrt{0.8992 \times 0.1008 \times 0.8951 \times 0.1049}} = 0.0553.
\end{aligned}$$

3) DL complet ($|\mathcal{D}'| = 1$, $|r| < 1$)

		$M2$		
		B	b	
$M1$	A	0.1029	0.0719	0.1748
	a	0.8252	0	0.8252
		0.9281	0.0719	1

$$\begin{aligned}
\mathcal{D}(AB) &= 0.1029 - 0.1748 \times 0.9281 = -0.0593, \\
\mathcal{D}'(AB) &= \frac{-0.0593}{\min(0.1748 \times 0.9281, 0.8252 \times 0.0719)} = -1, \\
r(AB) &= \frac{-0.0593}{\sqrt{0.1748 \times 0.8252 \times 0.9281 \times 0.0719}} = -0.6044.
\end{aligned}$$

4) DL total ($|\mathcal{D}'| = 1, |r| = 1$)

		$M2$		
		B	b	
$M1$	A	0.5763	0	0.5763
	a	0	0.4237	0.4237
		0.5763	0.4237	1

$$\begin{aligned}
\mathcal{D}(AB) &= 0.5763 - (0.5763)^2 = 0.2442, \\
\mathcal{D}'(AB) &= \frac{0.2442}{\min(0.1748 \times 0.9281, 0.8252 \times 0.0719)} = 1, \\
r(AB) &= \frac{0.2442}{\sqrt{0.1748 \times 0.8252 \times 0.9281 \times 0.0719}} = 1.
\end{aligned}$$

Les exemples 3 et 4 se distinguent quant à leur nombre d'haplotypes observés parmi les quatre possibles pour les deux marqueurs à l'étude, $M1$ et $M2$. On en observe trois dans l'exemple 3 et deux dans l'exemple 4. De plus, dans l'exemple 4, on remarque qu'un allèle particulier de $M1$ est toujours présent avec un allèle particulier de $M2$, d'où la corrélation de 1. Par exemple, l'allèle A n'a été observé qu'avec l'allèle B. Donc dans ce cas, si l'on sait que l'allèle A est présent au marqueur $M1$, on sait aussi que l'allèle B est présent au marqueur $M2$.

3.2 Stratégies d'analyse

Il existe différentes stratégies d'analyse pour l'étude d'association génétique. Deux devis d'échantillonnage sont possibles, l'échantillonnage de cas-témoins et l'échantillonnage de familles. De plus, deux approches analytiques peuvent être envisagées, l'approche allélique et l'approche haplotypique.

3.2.1 Devis d'échantillonnage

Échantillon de cas-témoins :

Deux échantillons indépendants sont recrutés. Un échantillon d'individus atteints de la maladie forme l'échantillon de cas. Un échantillon d'individus sains, non reliés aux malades, forme l'échantillon de témoins. Le chapitre 4 décrira la méthode d'analyse appropriée pour un échantillonnage de cas-témoins.

Échantillon de familles :

Puisque les études de cas-témoins sont sujettes à un problème de stratification de la population, comme on le verra à la section 4.3, un autre devis d'échantillonnage qui permet d'éviter ce problème a été proposé, à savoir l'échantillonnage de familles. Il consiste à recueillir l'information génétique d'un individu malade et des membres de sa famille (ses parents, ses frères et ses sœurs). Une famille peut présenter plus d'un cas de maladie. Très souvent, on utilise un échantillon de trios qui comprend l'individu atteint de la maladie et ses parents biologiques. Le chapitre 6 décrira la méthode d'analyse appropriée pour un échantillonnage de trios.

3.2.2 Approches analytiques

Étude d'association allélique :

On peut tester l'association entre une maladie et des marqueurs en étudiant un seul marqueur à la fois. On vérifie alors s'il y a association avec un allèle ou un génotype particulier.

Étude d'association haplotypique :

L'association entre une maladie et des marqueurs peut aussi être testée à partir des haplotypes que forment les marqueurs à l'étude. Cette approche permet de regrouper l'information de plusieurs marqueurs situés près l'un de l'autre et de tenir compte de l'éventuelle dépendance entre eux. On effectue d'abord un test global pour déterminer si au moins un des haplotypes est associé à la maladie. Si tel est le cas, on poursuit l'investigation pour identifier par des tests spécifiques le ou les haplotypes qui se révèlent plus fréquents chez les malades.

Si l'on connaît le ou les haplotypes potentiellement associés dans les études antérieures, on peut se limiter aux tests spécifiques de ces haplotypes sans avoir à corriger le seuil observé. Sinon, on teste spécifiquement chacun des haploytpes en prenant soin de corriger les seuils observés, avec une correction de Bonferroni par exemple.

Chapitre 4

Description de tests d'association dans un échantillon de cas-témoins

Il est possible d'éprouver la présence d'association allélique ou haplotypique au moyen de tests statistiques. Les procédures les plus classiques sont présentées ici. Dans chaque cas, l'implantation du test est illustrée au moyen des données provenant de l'échantillon fictif de cas-témoins présenté au tableau 4.1.

sujet		SNP 1	SNP 2	sujet		SNP 1	SNP 2
1	cas	C/C	A/A	16	témoin	C/T	A/G
2	témoin	C/T	A/G	17	témoin	C/C	A/G
3	témoin	C/C	A/A	18	cas	C/C	A/A
4	témoin	T/T	A/G	19	cas	C/C	A/A
5	témoin	C/C	A/A	20	témoin	C/T	G/G
6	témoin	T/T	G/G	21	témoin	T/T	G/G
7	cas	C/C	A/A	22	cas	C/C	A/A
8	témoin	C/C	A/A	23	cas	C/C	A/A
9	cas	C/C	A/A	24	cas	C/C	A/A
10	cas	C/C	A/A	25	cas	C/C	A/A
11	témoin	C/C	A/A	26	témoin	C/C	A/A
12	témoin	C/C	A/A	27	cas	C/C	A/A
13	cas	C/T	A/G	28	témoin	C/T	A/G
14	cas	C/C	A/A	29	cas	C/C	A/A
15	cas	C/C	A/A	30	témoin	T/T	G/G

TAB. 4.1 – Génotypes des individus d'un échantillon fictif de 15 cas et 15 témoins pour les SNP 1 et 2

4.1 Test d'association allélique

Considérons un marqueur à k allèles dénotés $a_1, \ldots, a_k$. Pour tester l'association allélique, on compare la distribution des allèles, ou encore celle des génotypes, chez les cas et les témoins. Étant donné un échantillon de taille n comportant n_1 cas et n_2 témoins, les données observées peuvent être présentées comme au tableau 4.2. Notons que le tableau compte au total $2n$ observations, puisque chaque individu possède deux allèles.

		cas	témoins	
allèle	a_1	n_{11}	n_{12}	
	$\vdots$	$\vdots$	$\vdots$	
	a_k	n_{k1}	n_{k2}	
		$2n_1$	$2n_2$	$2n$

TAB. 4.2 – Distribution des k allèles observés chez les cas et les témoins

Pour tester l'hypothèse

$$H_0 : \mathrm{P}(a_i|\text{cas}) = \mathrm{P}(a_i|\text{témoin}), \quad i = 1, \ldots, k.$$

on a généralement recours au test du khi-deux à $k-1$ degrés de liberté ; voir par exemple Clayton et Hills (1993). Si les distributions des deux groupes sont significativement différentes, le marqueur sera dit associé à la maladie. Des rapports de cotes permettent alors de caractériser l'association. La procédure `casecontrol` du module `genetics` du logiciel SAS permet d'effectuer ce test, soit à partir des allèles ou des génotypes.

4.1.1 Exemple de test sur les allèles

Testons l'association des allèles du SNP 1 avec la maladie.

En se référant au tableau 4.3, on trouve

$$X^2 = \frac{(29 - 23.5)^2}{23.5} + \frac{(18 - 23.5)^2}{23.5} + \frac{(1 - 6.5)^2}{6.5} + \frac{(12 - 6.5)^2}{6.5} = 11.88.$$

En comparant cette statistique à une loi du khi-deux à 1 degré de liberté, on constate que le seuil observé du test est de 0.0006, ce qui conduit au rejet de l'hypothèse nulle.

		cas	témoins	
allèle	C	29 (0.97)	18 (0.60)	47 (0.78)
	T	1 (0.03)	12 (0.40)	13 (0.22)
		30	30	60

TAB. 4.3 – Nombre observé (fréquence relative) des allèles observés pour le SNP 1 chez les cas et les témoins de l'exemple

On en déduit que l'allèle C est significativement plus fréquent chez les cas que chez les témoins.

Dans cet exemple, la cote de l'allèle C chez les cas est 19.33 fois plus élevée que celle chez les témoins. En effet, on trouve

$$\frac{\mathrm{P}(C|\text{cas})}{1-\mathrm{P}(C|\text{cas})} = \frac{29/30}{1-29/30} = 29$$

et

$$\frac{\mathrm{P}(C|\text{témoins})}{1-\mathrm{P}(C|\text{témoins})} = \frac{18/30}{1-18/30} = 1.5,$$

d'où $29/1.5 = 19.33$. Étant donné l'échantillonnage de cas-témoins, on exprime l'association à partir d'un rapport de cotes plutôt qu'un risque relatif. Si cependant l'allèle C avait été très rare dans la population, ce qui n'est pas le cas dans cet exemple puisqu'on observe une fréquence de 0.78 dans l'échantillon, le rapport de cotes de 19.33 pourrait s'interpréter comme un risque relatif.

4.1.2 Exemple de test sur les génotypes

Testons cette fois l'association des génotypes du SNP 1 avec la maladie. La procédure est identique à celle pour le test sur les allèles. Notons qu'ici le grand total est égal à n, la taille de l'échantillon, puisque chaque individu a un seul génotype.

À l'aide du tableau 4.4, on voit que

$$\begin{aligned} X^2 &= \frac{(14-10.5)^2}{10.5} + \frac{(7-10.5)^2}{10.5} + \frac{(1-2.5)^2}{2.5} \\ &\quad + \frac{(4-2.5)^2}{2.5} + \frac{(0-2)^2}{2} + \frac{(4-2)^2}{2} \\ &= 8.13. \end{aligned}$$

		cas	témoins	
génotype	C/C	14 (0.93)	7 (0.47)	21 (0.70)
	C/T	1 (0.07)	4 (0.27)	5 (0.17)
	T/T	0	4 (0.27)	4 (0.13)
		15	15	30

TAB. 4.4 – Nombre observé (fréquence relative) des génotypes observés pour le SNP 1 chez les cas et les témoins de l'exemple

Cette statistique obéit asymptotiquement à une loi du khi-deux à deux degrés de liberté sous l'hypothèse nulle. Puisque le seuil observé du test est 0.0171, on est porté à rejeter H_0. Les données suggèrent donc que le génotype C/C est plus fréquent chez les cas que chez les témoins. Le rapport de cotes est de 16 si les génotypes C/T et T/T sont regroupés dans la même catégorie.

4.2 Test d'association haplotypique

4.2.1 L'algorithme EM (Espérance-Maximisation)

4.2.1.1 Introduction à la problématique

Les techniques actuelles de génotypage ne permettent pas d'identifier directement les haplotypes d'un individu. On peut lire son génotype à plusieurs loci, mais on ne peut pas encore déterminer la séquence des allèles sur chaque chromosome.

Soit r le nombre d'haplotypes distincts. On rappelle qu'un individu possède deux haplotypes. L'un est reçu du père et l'autre de la mère ; ils peuvent être identiques ou différents. Le nombre total de paires d'haplotypes possibles est égal à la somme du nombre de paires d'haplotypes identiques $\binom{r}{1} = r$ et du nombre de paires d'haplotypes différents $\binom{r}{2} = r(r-1)/2$.

Reprenons l'exemple du tableau 4.1. Le SNP 1 a deux allèles, C et T. Le SNP 2 a deux allèles, A et G. Il y a donc $r = 2 \times 2 = 4$ haplotypes possibles, à savoir $h_1 = CA$, $h_2 = CG$, $h_3 = TA$ et $h_4 = TG$. Quant au nombre possible de paires d'haplotypes, il est donné par

$$\binom{r}{2} + \binom{r}{1} = 6 + 4 = 10.$$

	SNP 2		
SNP 1	A/A	A/G	G/G
C/C	CA, CA	CA, CG	CG, CG
C/T	CA, TA	CA, TG ou TA, CG	CG, TG
T/T	TA, TA	TA, TG	TG, TG

TAB. 4.5 – Les dix paires d'haplotypes possibles selon les génotypes aux SNP 1 et 2 de l'exemple

En observant le tableau 4.5, on remarque que l'identification des haplotypes se fait facilement pour les sujets homozygotes aux deux SNP (cellules (1,1), (1,3), (3,1) et (3,3)). Il en va de même pour ceux qui sont hétérozygotes pour un seul des deux SNP (cellules (1,2), (2,1), (2,3) et (3,2)). Cependant, l'identification demeure ambiguë pour les individus doublement hétérozygotes (C/T et A/G). De façon plus générale, dès qu'un sujet est hétérozygote pour au moins deux SNP, l'identification des haplotypes est difficile, surtout si le génotype des parents n'est pas disponible. Des méthodes statistiques ont donc été développées pour estimer les fréquences haplotypiques.

Pour obtenir les estimateurs du maximum de vraisemblance des fréquences haplotypiques, Excoffier et Slatkin (1995) ont proposé une adaptation de l'algorithme EM, formalisé par Dempster et coll. (1977).

4.2.1.2 Estimation par la méthode du maximum de vraisemblance

Précisons d'abord la terminologie utilisée pour décrire l'algorithme EM. Le *phénotype* représente l'information génétique que l'on peut observer, c'est-à-dire le génotype à plusieurs loci sans distinction des haplotypes. Le terme *paire d'haplotypes* réfère à une combinaison particulière d'haplotypes qui ne peut être observée directement. La figure 4.1 illustre cette distinction.

Le nombre c_j de paires d'haplotypes pouvant mener au phénotype j est fonction du nombre de loci hétérozygotes s_j :

$$c_j = \begin{cases} 2^{s_j-1}, & \text{si } s_j > 0\,; \\ 1, & \text{si } s_j = 0. \end{cases}$$

On peut également voir c_j comme une somme d'indicatrices,

$$c_j = \sum_{k=1}^{r} \sum_{\ell=k}^{r} 1_j(h_k, h_\ell)$$

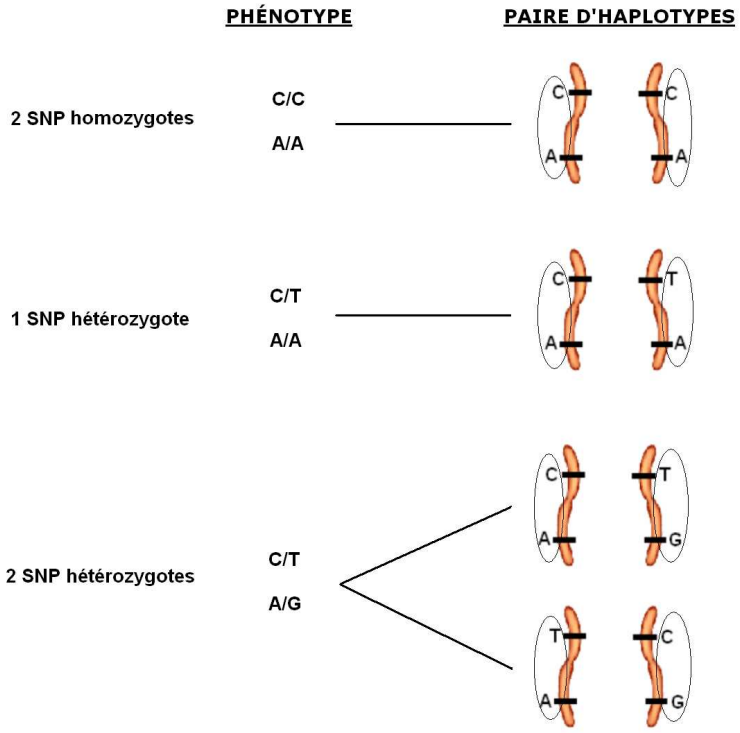


FIG. 4.1 – *Distinction entre phénotype et paire d'haplotypes*

où

$$1_j(h_k, h_\ell) = \begin{cases} 1, & \text{si les haplotypes } h_k, h_\ell \text{ permettent d'arriver au phénotype } j\,; \\ 0, & \text{sinon.} \end{cases}$$

Illustrons le principe de calcul de c_j à l'aide des données de l'exemple 4.1. Le tableau 4.6 en donne une compilation en fonction des neuf phénotypes possibles. Dans ce tableau, $j = 9$ correspond au phénotype C/T, A/G. Puisque les génotypes aux deux loci sont hétérozygotes, s_9 vaut 2. On a donc, pour la première formule,

$$c_9 = 2^{2-1} = 2.$$

Si l'on opte pour la deuxième formule, on a cette fois

$$\begin{aligned} c_9 &= 1_9(CA,CA) + 1_9(CA,CG) + 1_9(CA,TA) + 1_9(CA,TG) + 1_9(CG,CG) \\ &\quad + 1_9(CG,TA) + 1_9(CG,TG) + 1_9(TA,TA) + 1_9(TA,TG) + 1_9(TG,TG) \\ &= 0+0+0+1+0+1+0+0+0+0 \\ &= 2. \end{aligned}$$

j	phénotypes	n_j	s_j	c_j	paires d'haplotypes
1	C/C, A/A	20	0	1	CA, CA
2	C/C, G/G	0	0	1	CG, CG
3	T/T, A/A	0	0	1	TA, TA
4	T/T, G/G	3	0	1	TG, TG
5	C/T, A/A	0	1	1	CA, TA
6	C/T, G/G	1	1	1	CG, TG
7	C/C, A/G	1	1	1	CA, CG
8	T/T, A/G	1	1	1	TA, TG
9	C/T, A/G	4	2	2	CA, TG ou TA, CG
	total	30			

TAB. 4.6 – Phénotypes et paires d'haplotypes pour les 30 sujets de l'exemple

Soient n le nombre de sujets dans l'échantillon, m le nombre de phénotypes et n_j le nombre de sujets de l'échantillon ayant le phénotype j, où $j \in \{1,\ldots,m\}$. Le vecteur $(n_1,\ldots,n_m)$ obéit à une distribution multinomiale ayant pour paramètres le vecteur $(P_1,\ldots,P_m)$ des probabilités associées à chacun des phénotypes :

$$f(n_1,\ldots,n_m|P_1,\ldots,P_m) = \frac{n!}{n_1!\times\cdots\times n_m!}\, P_1^{n_1}\times\cdots\times P_m^{n_m}\ .$$

Les probabilités P_j s'obtiennent à partir de l'équation suivante :

$$P_j = \sum_{k=1}^{r}\sum_{\ell=k}^{r} 1_j(h_k,h_\ell)P(h_k,h_\ell), \tag{4.1}$$

où $P(h_k, h_\ell)$ représente la probabilité d'avoir la paire d'haplotypes h_k et h_ℓ ayant respectivement pour probabilités p_k et p_ℓ. En vertu du principe d'équilibre de Hardy–Weinberg,

$$P(h_k, h_\ell) = \begin{cases} {p_k}^2, & \text{si } k = \ell\,; \\ 2p_k p_\ell, & \text{si } k \neq \ell. \end{cases} \tag{4.2}$$

Ainsi, pour le phénotype C/C, A/A de l'exemple 4.1, on a

$$P_1 = 1_1(CA, CA)P(CA, CA) = p_1^2.$$

Pour le phénotype C/T, A/G, on a

$$\begin{aligned} P_9 &= 1_9(CA, TG)P(CA, TG) + 1_9(CG, TA)P(CG, TA) \\ &= 2p_1p_4 + 2p_2p_3. \end{aligned}$$

On voudrait estimer le vecteur $\theta = (p_1, p_2, p_3, p_4)$ des fréquences des quatre haplotypes ($h_1 = CA$, $h_2 = CG$, $h_3 = TA$ et $h_4 = TG$) par la méthode du maximum de vraisemblance. À une constante près, la vraisemblance est de la forme :

$$L(\theta|n_1, \ldots, n_m) = \prod_{j=1}^{m} P_j^{n_j} = \prod_{j=1}^{m} \left\{ \sum_{k=1}^{r} \sum_{\ell=k}^{r} 1_j(h_k, h_\ell) P(h_k, h_\ell) \right\}^{n_j}.$$

On cherche donc la valeur θ qui maximise :

$$\log L(\theta|n_1, \ldots, n_m) = \sum_{j=1}^{m} n_j \log\left(P_j\right). \tag{4.3}$$

Pour les données de l'exemple, le logarithme de la fonction de vraisemblance est le suivant (voir le tableau 4.6 pour les valeurs n_j) :

$$\begin{aligned} \log L(\theta|n_1, \ldots, n_9) &= n_1 \log\left({p_1}^2\right) + n_2 \log\left({p_2}^2\right) + n_3 \log\left({p_3}^2\right) + n_4 \log\left({p_4}^2\right) \\ &\quad + n_5 \log\left(2p_1p_3\right) + n_6 \log\left(2p_2p_4\right) + n_7 \log\left(2p_1p_2\right) \\ &\quad + n_8 \log\left(2p_3p_4\right) + n_9 \log\left(2p_1p_4 + 2p_2p_3\right) \\ &= 20 \log\left({p_1}^2\right) + 3 \log\left({p_4}^2\right) + \log\left(2p_2p_4\right) + \log\left(2p_1p_2\right) \\ &\quad + \log\left(2p_3p_4\right) + 4 \log\left(2p_1p_4 + 2p_2p_3\right). \end{aligned}$$

En principe, on obtiendrait $\hat{\theta} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$ par la résolution des équations obtenues en annulant les dérivées partielles. Afin de s'assurer que la contrainte $\hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 = 1$ soit respectée, on introduit un lagrangien λ en définissant :

$$g(p_1, \ldots, p_4, \lambda) = \log L - \lambda \left(\sum_{k=1}^{4} p_k - 1 \right).$$

Le système d'équations à résoudre est alors le suivant :

$$\begin{cases} \dfrac{\partial g}{\partial p_1} = \dfrac{2n_1 + n_5 + n_7}{p_1} + \dfrac{p_4 n_9}{p_1 p_4 + p_2 p_3} - \lambda = 0\,; \\ \dfrac{\partial g}{\partial p_2} = \dfrac{2n_2 + n_6 + n_7}{p_2} + \dfrac{p_3 n_9}{p_1 p_4 + p_2 p_3} - \lambda = 0\,; \\ \dfrac{\partial g}{\partial p_3} = \dfrac{2n_3 + n_5 + n_8}{p_3} + \dfrac{p_2 n_9}{p_1 p_4 + p_2 p_3} - \lambda = 0\,; \\ \dfrac{\partial g}{\partial p_4} = \dfrac{2n_4 + n_6 + n_8}{p_4} + \dfrac{p_1 n_9}{p_1 p_4 + p_2 p_3} - \lambda = 0\,; \\ \dfrac{\partial g}{\partial \lambda} = -\sum_{k=1}^{4} p_k + 1 = 0. \end{cases}$$

Cette procédure devient pénible lorsque le nombre d'haplotypes est grand. On peut alors avoir recours à l'algorithme EM afin de faciliter le calcul de ces estimations de fréquences haplotypiques.

4.2.1.3 Les étapes de l'algorithme EM

La solution proposée par Excoffier et Slatkin (1995) exploite l'algorithme EM, bien connu pour son efficacité dans la recherche d'estimations en présence de données incomplètes. Dans le cadre d'estimation des fréquences haplotypiques, on dispose effectivement de données incomplètes, car on a plus de catégories (les paires d'haplotypes) que ce qui est distinguable (les phénotypes). Si l'on connaissait la paire d'haplotypes de chaque sujet, on obtiendrait sans difficulté les estimateurs du maximum de vraisemblance pour les fréquences d'haplotypes.

Pour les données de l'exemple du tableau 4.1, l'estimation de la fréquence de l'haplotype $h_1 = CA$ s'obtiendrait comme suit, avec $k \neq 1$ et $\ell \neq 1$:

$$\begin{aligned} \hat{p}_1 &= \frac{\text{nombre d'haplotypes } h_1}{\text{nombre total d'haplotypes}} \\ &= \frac{2(\text{nombre de sujets } h_1h_1) + 1(\text{nombre de sujets } h_1h_k) + 0(\text{nombre de sujets } h_kh_\ell)}{2n} \\ &= \frac{2(n_1) + (n_5 + n_7 + \text{une partie de } n_9)}{2n} \\ &= \frac{2(20) + (0 + 1 + ?)}{60}. \end{aligned}$$

On voit ici que ce sont les $n_9 = 4$ sujets doublement hétérozygotes qui rendent l'estimation des fréquences haplotypiques ardue. Le nombre de sujets ayant la paire d'haplotypes CA, TG est manquant. Sachant que c'est une partie de n_9, on peut l'estimer par son espérance, mais cette estimation dépend des p_k :

$$\begin{aligned}
\text{E}(\#\text{ de sujets ayant les haplotypes } CA,\, TG) &= n_9 \times \left(\frac{P(CA,TG)}{P_9}\right) \\
&= n_9 \times \left(\frac{P(CA,TG)}{P(CA,TG)+P(TA,CG)}\right) \\
&= n_9 \times \left(\frac{2p_1p_4}{2p_1p_4+2p_2p_3}\right).
\end{aligned}$$

L'algorithme EM permet de contourner cette difficulté par un processus itératif à deux étapes : l'étape de *Maximisation* et l'étape d'*Espérance*.

Les valeurs initiales $\hat{p}_k^{(0)}$ des $\hat{p}_k$ doivent d'abord être choisies selon diverses options. On peut les générer aléatoirement ou encore choisir des fréquences équidistribuées. On peut aussi choisir des fréquences initiales égales au produit des fréquences d'allèles, ce qui suppose alors que les SNP sont en équilibre de liaison.

Étape de *maximisation*$^{(g)}$:

– on détermine les $\hat{p}_k^{(g)}$ à partir des $P_j^{(g-1)}$, à savoir

$$\hat{p}_k^{(g)} = \frac{1}{2}\sum_{j=1}^{m}\left[\sum_{i=1}^{c_j}\delta_{ijk}\frac{n_j}{n}\left\{\frac{P^{(g-1)}(\text{paire d'haplotypes } i)}{P_j^{(g-1)}}\right\}\right],$$

où δ_{ijk} est l'indicateur du nombre de fois (0, 1 ou 2) qu'apparaît l'haplotype h_k dans la paire d'haplotypes i du phénotype j et $P^{(g-1)}(\text{paire d'haplotypes } i)$ est défini à l'équation (4.2) ;

– notons que $P^{(g-1)}(\text{paire d'haplotypes } i)/P_j^{(g-1)} = 1$ lorsque $c_j = 1$.

Étape d'*espérance*$^{(g)}$:

– on détermine les $P_j^{(g)}$ à l'aide de l'équation (4.1) :

$$P_j^{(g)} = \sum_{k=1}^{r}\sum_{\ell=k}^{r} 1_j(h_k,h_\ell)P(h_k,h_\ell),$$

où $P^{(g-1)}(h_k,h_\ell)$ est défini à l'équation (4.2) et

$$1_j(h_k,h_\ell) = \begin{cases} 1, & \text{si les haplotypes } h_k, h_\ell \text{ permettent d'arriver au phénotype } j\,; \\ 0, & \text{sinon.} \end{cases}$$

– on calcule le logarithme de la fonction de vraisemblance :

$$\log L^{(g)} = \sum_{j=1}^{m} n_j \log\left(P_j^{(g)}\right).$$

Notons que l'indice supérieur (g) réfère à la g-ième itération de l'algorithme. On en effectue plusieurs jusqu'à ce que l'algorithme converge. Le critère de convergence suivant est couramment utilisé :

$$\text{ratio différence} = \frac{\log L^{(g)} - \log L^{(g+1)}}{\log L^{(g)}} < 10^{-5}.$$

Afin d'éviter la convergence vers un maximum local plutôt que vers le maximum global, il est recommandé de répéter l'algorithme EM avec différentes valeurs initiales.

La procédure `haplotype` de SAS et la fonction `haplo.em` du logiciel R permettent d'obtenir les estimations des fréquences haplotypiques via l'algorithme EM.

4.2.1.4 Illustration des étapes de l'algorithme EM

Le tableau 4.7 illustre la convergence de l'algorithme pour les données du tableau 4.1. Les fréquences haplotypiques $\hat{p}_1, \hat{p}_2, \hat{p}_3$ et $\hat{p}_4$ correspondent respectivement aux fréquences des haplotypes CA, CG, TA et TG. Des valeurs initiales équidistribuées ainsi que des valeurs basées sur le produit des fréquences d'allèles ont été utilisées.

Afin d'illustrer le fonctionnement de l'algorithme EM, le détail des calculs correspondant aux trois premières itérations est présenté dans le cas où les valeurs initiales sont basées sur le produit des fréquences d'allèles.

Itération 0

– Initialisation des fréquences haplotypiques :

$$\begin{aligned}
\hat{p}_1^{(0)} &= p(\text{allèle } C \text{ au SNP 1})p(\text{allèle } A \text{ au SNP 2}) \\
&= 0.78\overline{33} \times 0.76\overline{66} \\
&= 0.60056
\end{aligned}$$

$$\begin{aligned}
\hat{p}_2^{(0)} &= p(\text{allèle } C \text{ au SNP 1})p(\text{allèle } G \text{ au SNP 2}) \\
&= 0.78\overline{33} \times 0.23\overline{33} \\
&= 0.18278
\end{aligned}$$

itération (g)	$\log L$	ratio différence	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$
			valeurs initiales équidistribuées			
0	−75.55304		0.25000	0.25000	0.25000	0.25000
1	−39.94232	0.47133	0.71667	0.06667	0.05000	0.16667
2	−38.29515	0.04124	0.74819	0.03514	0.01848	0.19819
3	−38.28870	0.00017	0.74971	0.03362	0.01696	0.19971
4	−38.28869	0.00000	0.74975	0.03359	0.01692	0.19975
			valeurs initiales = produit fréq. allèles			
0	−56.33114		0.60056	0.18278	0.16611	0.05056
1	−39.94232	0.29094	0.71667	0.06667	0.05000	0.16667
2	−38.29515	0.04124	0.74819	0.03514	0.01848	0.19819
3	−38.28870	0.00017	0.74971	0.03362	0.01696	0.19971
4	−38.28869	0.00000	0.74975	0.03359	0.01692	0.19975

TAB. 4.7 – Itérations de l'algorithme EM pour l'estimation des fréquences haplotypiques de l'exemple

$$\begin{aligned}\hat{p}_3^{(0)} &= p(\text{allèle } T \text{ au SNP 1})p(\text{allèle } A \text{ au SNP 2}) \\ &= 0.21\overline{66} \times 0.76\overline{66} \\ &= 0.16611\end{aligned}$$

$$\begin{aligned}\hat{p}_4^{(0)} &= p(\text{allèle } T \text{ au SNP 1})p(\text{allèle } G \text{ au SNP 2}) \\ &= 0.21\overline{66} \times 0.23\overline{33} \\ &= 0.05056\end{aligned}$$

– Étape d'*espérance*$^{(0)}$: on calcule $\log L^{(0)}$, soit

$$\begin{aligned}\log L^{(0)} &= 20 \log\left\{(0.60056)^2\right\} + 3 \log\left\{(0.05056)^2\right\} \\ &\quad + \log(2 \times 0.18278 \times 0.05056) \\ &\quad + \log(2 \times 0.60056 \times 0.18278) + \log(2 \times 0.16611 \times 0.05056) \\ &\quad + 4 \log(2 \times 0.60056 \times 0.05056 + 2 \times 0.18278 \times 0.16611) \\ &= -56.33114.\end{aligned}$$

ITÉRATION 1

– Étape de *maximisation*$^{(1)}$: on maximise les fréquences haplotypiques, ce qui revient dans ce cas-ci à maximiser le nombre de sujets ayant la paire d'haplotypes

CA, TG et le nombre de sujets ayant la paire d'haplotypes TA, CG :

$$\begin{aligned}
\hat{p}_1^{(1)} &= \frac{1}{2}\left[2 \times \frac{20}{30}\left\{\frac{(\hat{p}_1^{(0)})^2}{(\hat{p}_1^{(0)})^2}\right\} + \frac{1}{30}\left\{\frac{2\hat{p}_1^{(0)}\hat{p}_2^{(0)}}{2\hat{p}_1^{(0)}\hat{p}_2^{(0)}}\right\}\right] \\
&\quad + \frac{1}{2}\left\{\frac{4}{30}\left(\frac{2\hat{p}_1^{(0)}\hat{p}_4^{(0)}}{2\hat{p}_1^{(0)}\hat{p}_4^{(0)} + 2\hat{p}_2^{(0)}\hat{p}_3^{(0)}}\right)\right\} \\
&= \frac{1}{2}\left\{\frac{41}{30} + \frac{4}{30}\left(\frac{2 \times 0.60056 \times 0.05056}{2 \times 0.60056 \times 0.05056 + 2 \times 0.18278 \times 0.16611}\right)\right\} \\
&= \frac{1}{2}\left(\frac{41}{30} + \frac{\mathbf{2}}{30}\right) \\
&= \frac{43}{60} = 0.716667 \\
\hat{p}_2^{(1)} &= \frac{1}{2}\left\{\frac{2}{30} + \frac{4}{30}\left(\frac{2 \times 0.18278 \times 0.16611}{2 \times 0.60056 \times 0.05056 + 2 \times 0.18278 \times 0.16611}\right)\right\} \\
&= \frac{1}{2}\left(\frac{2}{30} + \frac{\mathbf{2}}{30}\right) \\
&= \frac{4}{60} = 0.066667 \\
\hat{p}_3^{(1)} &= \frac{1}{2}\left\{\frac{1}{30} + \frac{4}{30}\left(\frac{2 \times 0.18278 \times 0.16611}{2 \times 0.60056 \times 0.05056 + 2 \times 0.18278 \times 0.16611}\right)\right\} \\
&= \frac{1}{2}\left(\frac{1}{30} + \frac{\mathbf{2}}{30}\right) \\
&= \frac{3}{60} = 0.05 \\
\hat{p}_4^{(1)} &= \frac{1}{2}\left\{\frac{8}{30} + \frac{4}{30}\left(\frac{2 \times 0.60056 \times 0.05056}{2 \times 0.60056 \times 0.05056 + 2 \times 0.18278 \times 0.16611}\right)\right\} \\
&= \frac{1}{2}\left(\frac{8}{30} + \frac{\mathbf{2}}{30}\right) \\
&= \frac{10}{60} = 0.166667
\end{aligned}$$

– Étape d'*espérance*$^{(1)}$: on calcule $\log L^{(1)}$:

$$\begin{aligned}
\log L^{(1)} &= 20 \log\left\{(0.716667)^2\right\} + 3 \log\left\{(0.166667)^2\right\} \\
&\quad + \log\left(2 \times 0.066667 \times 0.166667\right) \\
&\quad + \log\left(2 \times 0.716667 \times 0.066667\right) + \log\left(2 \times 0.05 \times 0.166667\right) \\
&\quad + 4 \log\left(2 \times 0.716667 \times 0.166667 + 2 \times 0.066667 \times 0.05\right) \\
&= -39.94232.
\end{aligned}$$

ITÉRATION 2

– Étape de *maximisation*$^{(2)}$:

$$\begin{aligned}
\hat{p}_1^{(2)} &= \frac{1}{2}\left\{\frac{41}{30}+\frac{4}{30}\left(\frac{2\times 0.71667\times 0.16667}{2\times 0.71667\times 0.16667+2\times 0.06667\times 0.05}\right)\right\}\\
&= \frac{1}{2}\left(\frac{41}{30}+\frac{\mathbf{3.8914}}{30}\right)\\
&= 0.74819
\end{aligned}$$

$$\begin{aligned}
\hat{p}_2^{(2)} &= \frac{1}{2}\left\{\frac{2}{30}+\frac{4}{30}\left(\frac{2\times 0.06667\times 0.05}{2\times 0.71667\times 0.16667+2\times 0.06667\times 0.05}\right)\right\}\\
&= \frac{1}{2}\left(\frac{2}{30}+\frac{\mathbf{0.1086}}{30}\right)\\
&= 0.03514
\end{aligned}$$

$$\begin{aligned}
\hat{p}_3^{(2)} &= \frac{1}{2}\left\{\frac{1}{30}+\frac{4}{30}\left(\frac{2\times 0.06667\times 0.05}{2\times 0.71667\times 0.16667+2\times 0.06667\times 0.05}\right)\right\}\\
&= \frac{1}{2}\left(\frac{1}{30}+\frac{\mathbf{0.1086}}{30}\right)\\
&= 0.01848
\end{aligned}$$

$$\begin{aligned}
\hat{p}_4^{(2)} &= \frac{1}{2}\left\{\frac{8}{30}+\frac{4}{30}\left(\frac{2\times 0.71667\times 0.16667}{2\times 0.71667\times 0.16667+2\times 0.06667\times 0.05}\right)\right\}\\
&= \frac{1}{2}\left(\frac{8}{30}+\frac{\mathbf{3.8914}}{30}\right)\\
&= 0.19819
\end{aligned}$$

– Étape d'*espérance*$^{(2)}$: on calcule $\log L^{(2)}$:

$$\begin{aligned}
\log L^{(2)} &= 20\log\left\{(0.74819)^2\right\}+3\log\left\{(0.19819)^2\right\}\\
&\quad +\log\left(2\times 0.03514\times 0.19819\right)\\
&\quad +\log\left(2\times 0.74819\times 0.03514\right)+\log\left(2\times 0.01848\times 0.19819\right)\\
&\quad +4\log\left(2\times 0.74819\times 0.19819+2\times 0.03514\times 0.01848\right)\\
&= -38.29515.
\end{aligned}$$

4.2.2 Le test du rapport des vraisemblances

Pour détecter la présence d'association entre la maladie et les haplotypes, un test du rapport des vraisemblances a été proposé par Zhao et coll. (2000). Ce test permet de comparer les fréquences des haplotypes des cas et des témoins. L'algorithme EM est utilisé pour obtenir les estimations des fréquences haplotypiques. On se rappelle que la procédure `haplotype` de SAS pouvait alors être utilisée. Cette procédure permet aussi d'effectuer le test du rapport des vraisemblances proposé par Zhao et coll. (2000).

4.2.2.1 Retour à l'exemple

Le tableau 4.8 présente la distribution des phénotypes chez les cas et chez les témoins de l'exemple. À la section 4.2.1, les fréquences haplotypiques des 30 sujets, c'est-à-dire l'échantillon combiné, ont été estimées par l'algorithme EM. Il est aussi possible d'obtenir des estimations pour les cas et les témoins séparément ; voir les tableaux 4.9 et 4.10.

		nombre de sujets		
j	phénotypes	cas	témoins	combiné
1	C/C, A/A	14	6	20
2	C/C, G/G	0	0	0
3	T/T, A/A	0	0	0
4	T/T, G/G	0	3	3
5	C/T, A/A	0	0	0
6	C/T, G/G	0	1	1
7	C/C, A/G	0	1	1
8	T/T, A/G	0	1	1
9	C/T, A/G	1	3	4
		15	15	30

TAB. 4.8 – Distribution des phénotypes chez les cas et les témoins de l'exemple

	fréquences haplotypiques		
	cas	témoins	combiné
$\hat{p}_1$	0.96667	0.53213	0.74975
$\hat{p}_2$	0	0.06787	0.03359
$\hat{p}_3$	0	0.03454	0.01692
$\hat{p}_4$	0.03333	0.36546	0.19975

TAB. 4.9 – Estimation des fréquences haplotypiques chez les cas et les témoins de l'exemple

	nombre d'haplotypes		
	cas	témoins	combiné
CA	29	15.96	44.96
CG	0	2.04	2.04
TA	0	1.04	1.04
TG	1	10.96	11.96
	30	30	60

TAB. 4.10 – Nombre d'haplotypes estimé chez les cas et les témoins de l'exemple

4.2.2.2 Test global

On teste l'hypothèse

$$H_0 : \text{aucun haplotype n'est associé à la maladie.}$$

La statistique du test du rapport des vraisemblances est :

$$\Lambda = 2\left(\log L_{\text{cas}} + \log L_{\text{témoins}} - \log L_{\text{combiné}}\right).$$

La loi asymptotique de la statistique Λ est khi-deux sous H_0. Les degrés de liberté pour les cas, les témoins et l'échantillon combiné doivent être calculés séparément. Ils correspondent au nombre d'haplotypes distincts observés dans chacun des groupes moins un. Pour le nombre de degrés de liberté de la statistique Λ, on a donc

$$\nu = \nu_{\text{cas}} + \nu_{\text{témoins}} - \nu_{\text{combiné}}.$$

Notons que si tous les haplotypes possibles sont observés chez les cas et chez les témoins, le nombre de degrés de liberté sera simplement égal au nombre d'haplotypes moins un.

Pour l'exemple du tableau 4.1, les logarithmes de la fonction de vraisemblance, qui s'obtiennent à partir de l'équation (4.3) et des données des tableaux 4.8 et 4.9, sont :

$$\begin{aligned}
\log L_{\text{cas}} &= 14 \log\left\{(0.96667)^2\right\} + \log\left(2 \times 0.96667 \times 0.03333 + 0\right) = -3.6912, \\
\log L_{\text{témoins}} &= 6 \log\left\{(0.53213)^2\right\} + 3 \log\left\{(0.36546)^2\right\} \\
&\quad + \log\left(2 \times 0.06787 \times 0.36546\right) \\
&\quad + \log\left(2 \times 0.53213 \times 0.06787\right) + \log\left(2 \times 0.03454 \times 0.36546\right) \\
&\quad + 3 \log\left(2 \times 0.53213 \times 0.36546 + 2 \times 0.06787 \times 0.03454\right) = -25.7176
\end{aligned}$$

et

$$\log L_{\text{combiné}} = -38.2887 \text{(voir le tableau 4.7)}.$$

La statistique du test est $\Lambda = 2\left(-3.6912 - 25.7176 + 38.2887\right) = 17.7598$. En la comparant à une loi du khi-deux à trois degrés de liberté, on estime le seuil du test à 0.0005. On rejette donc l'hypothèse nulle et on conclut qu'au moins un des haplotypes est associé à la maladie.

4.2.2.3 Tests spécifiques

On teste cette fois une hypothèse plus spécifique, soit

$$H_0 : \text{l'haplotype } j \text{ n'est pas associé à la maladie.}$$

Disons que l'on veut tester l'association entre la maladie et l'haplotype CA de l'exemple. On crée alors à partir du tableau 4.10 un tableau qui regroupe les trois autres haplotypes dans la même catégorie tel qu'illustré au tableau 4.11.

La statistique du test est la statistique du khi-deux habituelle, à savoir :

$$X^2 = \frac{(29 - 22.48)^2}{22.48} + \frac{(15.96 - 22.48)^2}{22.48} + \frac{(1 - 7.52)^2}{7.52} + \frac{(14.04 - 7.52)^2}{7.52} = 15.09.$$

On procède de la même façon pour les haplotypes CG, TA et TG. Le tableau 4.12 présente les résultats des tests spécifiques pour l'exemple.

	nombre d'haplotypes		
	cas	témoins	
haplotype CA	29	15.96	44.96
les 3 autres	1	14.04	15.04
	30	30	60

TAB. 4.11 – Tableau permettant de tester spécifiquement l'association de l'haplotype CA

	test spécifique	
haplotype	X^2	seuil observé
CA	15.09	0.0001
CG	2.13	0.1445
TA	1.08	0.2995
TG	10.35	0.0013

TAB. 4.12 – Résultats des tests d'association spécifiques pour l'exemple

Au vu des résultats, on conclut qu'au seuil de 5%, l'haplotype CA est significativement plus fréquent chez les gens malades que chez les gens sains, alors que l'haplotype TG l'est moins. Ces observations demeurent vraies même si l'on utilise une correction de Bonferonni pour les quatre tests puisqu'alors les seuils observés sont comparés à la valeur 1.25%.

4.2.3 Le test du score

Un autre test d'association haplotypique a été proposé par Schaid et coll. (2002). Celui-ci s'appuie sur une statistique du score basée sur un modèle linéaire généralisé. Pour des fins de simplification, on ne présente ici que la procédure de base permettant de tester l'association des haplotypes avec un trait dichotomique (0 : sain, 1 : malade) et sans covariables environnementales. Le lecteur intéressé à connaître la forme plus générale du test est encouragé à consulter l'article de Schaid et coll. (2002). Mentionnons que la fonction `haplo.score` du logiciel R permet d'effectuer ce type de test.

4.2.3.1 Contexte

Soit un échantillon de cas-témoins de taille n. Supposons que le nombre d'haplotypes possibles pour les marqueurs à l'étude soit r. Posons y_i, le trait de l'individu i, et $\mathbf{X}_i$, le vecteur des codes numériques pour la paire d'haplotypes de l'individu i.

Le vecteur $\mathbf{X}_i$ est une fonction quelconque de la paire d'haplotypes. On choisit souvent un vecteur de dimension égale au nombre d'haplotypes. Les coefficients correspondent au nombre d'haplotypes de chaque type que l'individu i possède (0, 1 ou 2) selon la paire d'haplotypes qu'il a. Par exemple, pour les données du tableau 4.1, le premier élément du vecteur représentera le nombre d'haplotypes CA que possède le sujet i, le deuxième correspondra au nombre d'haplotypes CG et de même pour les deux autres éléments. Pour un sujet ayant la paire d'haplotypes CA, CA, on aura $X_i^\top = (2, 0, 0, 0)$.

4.2.3.2 Test global

On teste l'hypothèse

$$H_0 : \text{aucun haplotype n'est associé à la maladie.}$$

La statistique du test du score est la suivante :

$$S = U^\top V^{-1} U,$$

où U est un vecteur colonne dont la dimension r représente le nombre d'haplotypes distincts. Cette fonction score U se définit comme suit :

$$U = \sum_{i=1}^{n} (y_i - \overline{y}) \mathrm{E}(\mathbf{X}_i).$$

Ici, $\overline{y}$ représente la proportion de gens atteints de la maladie dans l'échantillon étudié.

La matrice de variance-covariance V, de dimension $r \times r$, est donnée par

$$\begin{aligned} V = & \sum_{i=1}^{n} (y_i - \overline{y})^2 \mathrm{E}(\mathbf{X}_i)\mathrm{E}(\mathbf{X}_i^\top) \\ & - \left\{ \sum_{i=1}^{n} \overline{y}(1-\overline{y}) \mathrm{E}(\mathbf{X}_i^\top) \right\}^\top \left\{ \sum_{i=1}^{n} \overline{y}(1-\overline{y}) \right\}^{-1} \left\{ \sum_{i=1}^{n} \overline{y}(1-\overline{y}) \mathrm{E}(\mathbf{X}_i^\top) \right\}. \end{aligned}$$

Dans les expressions pour U et V, $\mathrm{E}(\mathbf{X}_i)$ représente la paire d'haplotypes espérée pour l'individu i sachant son phénotype. On l'obtient à partir de l'équation suivante, en sommant sur toutes les paires d'haplotypes possibles :

$$\mathrm{E}(\mathbf{X}_i) = \textstyle\sum_{g \in G} X(g) Q(g).$$

Ici, G représente l'ensemble de toutes les paires d'haplotypes possibles et $Q(g)$ correspond à la probabilité d'observer la paire d'haplotypes g selon le phénotype de l'individu.

À l'aide de l'exemple, illustrons le principe du calcul de $\mathrm{E}(\mathbf{X}_i)$ quand l'identification des haplotypes nécessite l'utilisation de l'algorithme EM. Prenons le sujet n$^{\mathrm{o}}$ 2 dont le phénotype est C/T, A/G. Deux combinaisons d'haplotypes sont possibles : CA, TG ou CG, TA. Les estimations des fréquences haplotypiques calculées avec l'algorithme EM nous permettent d'estimer la probabilité de chacune des possibilités :

- $\mathrm{P}\,(\text{haplotypes } CA,TG \mid \text{phénotype } C/T,A/G) = \dfrac{2p_1p_4}{2p_1p_4 + 2p_2p_3} = 0.9962\,;$
- $\mathrm{P}\,(\text{haplotypes } CG,TA \mid \text{phénotype } C/T,A/G) = \dfrac{2p_2p_3}{2p_1p_4 + 2p_2p_3} = 0.0038.$

On a donc

$$\mathrm{E}(\mathbf{X}_2) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} 0.9962 + \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} 0.0038 = \begin{pmatrix} 0.9962 \\ 0.0038 \\ 0.0038 \\ 0.9962 \end{pmatrix}.$$

Si les haplotypes d'un individu sont facilement identifiables sans l'aide de l'algorithme EM, $\mathrm{E}(\mathbf{X}_i)$ sera simplement égal à $\mathbf{X}_i$.

Pour les données de l'exemple, la fonction score et la matrice de variance-covariance sont respectivement données par

$$U = \begin{pmatrix} 6.5038 \\ -1.0038 \\ -0.5038 \\ -4.9962 \end{pmatrix} \quad \text{et} \quad V = \begin{pmatrix} 4.38 & -0.5 & -0.38 & -3.5 \\ -0.5 & 0.47 & -0.02 & 0.05 \\ -0.38 & -0.02 & 0.24 & 0.15 \\ -3.5 & 0.05 & 0.15 & 3.3 \end{pmatrix}.$$

La statistique du test vaut $S = 9.82531$ et peut être comparée, sous H_0, à une loi du khi-deux à trois degrés de liberté. L'approximation asymptotique du seuil vaut 0.0201. Au seuil de 5%, on conclut donc qu'il existe une association entre la maladie et au moins un des haplotypes.

4.2.3.3 Tests spécifiques

Pour tester l'hypothèse

$$H_0 : \text{l'haplotype } j \text{ n'est pas associé à la maladie,}$$

on calcule la statistique du score suivante :

$$Z_j = \frac{U[j]}{\sqrt{V[j,j]}} \quad \text{pour } j = 1, \ldots, r$$

Sous H_0, Z_j obéit asymptotiquement à une loi normale centrée réduite.

Les résultats pour les tests spécifiques de l'exemple sont :

$$\begin{aligned}
Z_{CA} &= \frac{U[1]}{\sqrt{V[1,1]}} = \frac{6.5038}{\sqrt{4.38}} = 3.108, \\
Z_{CG} &= \frac{U[2]}{\sqrt{V[2,2]}} = \frac{-1.0038}{\sqrt{0.47}} = -1.464, \\
Z_{TA} &= \frac{U[3]}{\sqrt{V[3,3]}} = \frac{-0.5038}{\sqrt{0.24}} = -1.028, \\
Z_{TG} &= \frac{U[4]}{\sqrt{V[4,4]}} = \frac{-4.9962}{\sqrt{3.3}} = -2.75.
\end{aligned}$$

Les seuils observés sont respectivement 0.0019, 0.1415, 0.3052 et 0.0059. Il semble donc qu'au seuil de 5% l'haplotype CA soit positivement associé à la maladie, alors que l'haplotype TG le serait négativement. Ces observations demeurent vraies même si l'on utilise une correction de Bonferroni pour les quatre tests puisqu'alors les seuils observés sont comparés à la valeur 1.25%.

4.3 Problème de stratification dans la population

Tel que mentionné au début du chapitre 3, il est possible que l'association que l'on détecte ne soit pas réelle. Une fausse association peut survenir s'il y a stratification dans la population, c'est-à-dire si les cas et les témoins ne sont pas homogènes. L'interprétation des résultats risque alors d'être erronée.

Un exemple fictif tiré du cours STT-66943 *Introduction à la statistique génétique*, donné par la professeure Aurélie Labbe à l'automne 2005, illustre bien le problème de stratification dans une population. Cet exemple concerne une étude sur la fibrose kystique réalisée à Vancouver (Colombie-Britannique), où l'on considère qu'il y a autant de caucasiens que d'asiatiques. Cet exemple est présenté ci-dessous.

4.3.1 Exemple de stratification dans une population

Le risque d'être d'atteint de fibrose kystique dépend de l'ethnicité. Cette maladie touche surtout les gens d'origine caucasienne. La prévalence est de 1 cas pour 3300 individus chez les caucasiens alors qu'elle est de 1 cas pour 32 100 chez les asiatiques (Wilson, 2002).

Un échantillon de 100 cas et de 100 témoins est recruté à Vancouver. On étudie un marqueur à deux allèles, B et b. Les fréquences d'allèles à ce marqueur observées dans l'échantillon sont présentées au tableau 4.13.

		nombre d'allèles		
		cas	témoins	
allèle	B	118	82	200
	b	82	118	200
		200	200	400

TAB. 4.13 – Distribution des allèles observés chez les cas et les témoins de l'échantillon de Vancouver

Les fréquences d'allèles espérées sont toutes égales à 100. Un test d'association allélique, tel que décrit à la section 4.1, détecte une très forte association entre l'allèle B et la maladie. La statistique du test du khi-deux vaut :

$$X^2 = \frac{(118-100)^2}{100} + \frac{(82-100)^2}{100} + \frac{(118-100)^2}{100} + \frac{(82-100)^2}{100} = 12.96.$$

Si l'on tient compte cette fois de l'origine ethnique des participants, les résultats sont très différents. En effet, les fréquences alléliques de la population de Vancouver pour ce marqueur sont connues ; elles dépendent de l'origine ethnique. Chez les caucasiens, la fréquence de l'allèle B est de 0.6 alors que, chez les asiatiques, elle est de 0.2. Étant donné la prévalence plus forte chez les caucasiens, supposons que les 100 cas sélectionnés sont tous de cette origine. Pour ce qui est des témoins, on suppose que la moitié d'entre eux

sont caucasiens et que l'autre moitié est asiatique. Les fréquences d'allèles espérées sont maintenant différentes de 100. Pour les cas, le nombre espéré d'allèles B est $200 \times 0.6 = 120$. Pour les témoins, il est de $(200 \times 0.6)/2 + (200 \times 0.2)/2 = 80$.

La statistique du test devient :

$$X^2 = \frac{(118-120)^2}{120} + \frac{(82-80)^2}{80} + \frac{(118-120)^2}{120} + \frac{(82-80)^2}{80} = 0.17.$$

On ne détecte maintenant plus d'association entre ce marqueur et la fibrose kystique. Le résultat précédant n'est pas valide, puisqu'il ne tient pas compte de la présence de stratification dans la population étudiée.

Il est donc important de porter une attention particulière lors de la sélection des témoins. On conseille généralement d'apparier les cas et les témoins selon l'âge, le sexe, la région géographique et l'origine ethnique.

Cette problématique liée à la stratification a motivé la recherche d'un devis d'échantillonnage qui n'aurait pas ce désavantage. Dans cette optique, Falk et Rubenstein (1987) ont proposé un devis de trios où les allèles transmis des parents à un enfant atteint seront les allèles *cas* alors que les allèles non transmis seront les allèles *témoins*. Le devis ainsi que les méthodes d'analyse qui y correspondent sont décrits au chapitre 6. Avant, on présente au prochain chapitre les résultats d'études d'association génétique pour la schizophrénie qui ont été réalisées dans des échantillons de cas-témoins.

Chapitre 5

Résultats d'études d'association génétique pour la schizophrénie

Le présent chapitre propose un survol des conclusions de quelques études d'association effectuées dans des échantillons de cas-témoins. Ces résultats ont été soit rapportés dans la littérature, soit obtenus dans le cadre du programme de recherche en psychiatrie génétique du Centre de recherche Université Laval Robert-Giffard (CRULRG).

La schizophrénie (SZ) est une maladie psychiatrique complexe caractérisée entre autres par des symptômes de délire et d'hallucination. On estime à 1% la prévalence de cette maladie dans la population. Les recherches menées jusqu'à présent n'ont pas permis d'identifier les gènes responsables de la SZ, mais quelques gènes candidats ont été repérés.

Parmi les gènes qui pourraient être associés à la SZ, on retrouve le gène dysbindin (DTNBP1), situé sur le chromosome 6, ainsi que le gène neuregulin (NRG1), situé sur le chromosome 8. Dernièrement, des études ont rapporté des associations entre ces gènes et la SZ. La première partie de ce chapitre résume deux d'entre elles. Au CRULRG, on s'intéresse aussi à l'identification des gènes de vulnérabilité à la SZ. Dans le cadre de ce mémoire, une étude d'association entre cette maladie et les deux gènes susmentionnés a été réalisée. Les résultats de cette étude font l'objet de la deuxième partie du présent chapitre.

5.1 Résultats d'études récentes

Le tableau 5.1 résume deux études d'association pour la SZ divulguées récemment. La première étude s'est intéressée au gène NRG1 alors que la seconde portait sur le gène DTNBP1.

auteurs	population étudiée	échantillon	gène	SNP	allèle associé[a]
Stefansson et al. (2003)	Écossais	cas-témoins	NRG1	nrg221132	-
				nrg221533	*C*
				nrg241930	*G*
				nrg243177	*T*
				nrg433E1006	-
Funke et al. (2004)	Américains	cas-témoins	DTNBP1	rs909706	-
				rs1018381	*T*
				rs2619522	-
				rs760761	-
				rs2619528	*A*
				rs1011313	-

[a] Allèle significativement plus fréquent chez les cas que chez les témoins.

Tab. 5.1 – Résumé de deux récentes études d'association pour la schizophrénie

En 2003, Stefansson et coll. (2003) ont publié les résultats d'une étude d'association pour la SZ réalisée auprès de la population écossaise. Ils disposaient d'un échantillon de 609 cas et 618 témoins. Ils ont évalué l'association de cinq SNP situés dans la région du gène NRG1. Les associations allélique et haplotypique des SNP ont été testées. Les fréquences des haplotypes ont été estimées par l'entremise de l'algorithme EM.

Dans leurs travaux, Stefansson et coll. (2003) ont détecté une association allélique significative pour trois des SNP à l'étude. L'allèle *C* du SNP nrg221533 s'est avéré plus fréquent chez les cas que chez les témoins ; le risque relatif a été estimé à 1.4 et le seuil observé est de l'ordre de 6.4×10^{-5}. Pour le gène nrg241930, il appert que l'allèle *G* est plus fréquent chez les sujets malades ; le risque relatif est de 1.3 et le seuil observé d'environ 0.0021. Avec un risque relatif de 1.3 et un seuil observé de 0.0008, l'allèle *T* du gène nrg243177 fait également partie des allèles qui semblent associés à la SZ.

De plus, Stefansson et coll. (2003) signalent que dans leur étude, l'haplotype $G\mathbf{CGT}G$ s'est révélé significativement plus fréquent chez les cas que chez les témoins ; le risque relatif est estimé à 1.5 et le seuil observé est très faible (3.2×10^{-5}).

Pour leur part, Funke et coll. (2004) ont publié les résultats d'une étude d'association réalisée au sein de la population américaine. Ils ont sélectionné des sujets provenant de trois origines ethniques, soit des caucasiens, des hispaniques et des afro-américains. Leur échantillon de caucasiens était composé de 258 cas et 467 témoins. Ces auteurs ont étudié l'association de six SNP situés dans la région du gène DTNBP1. Les associations allélique et haplotypique des SNP ont été testées. Des tests du khi-deux ont été effectués pour détecter l'association allélique ; de même, des tests du score ont été employés pour vérifier la présence d'association haplotypique. L'algorithme EM a été utilisé pour estimer les fréquences haplotypiques.

D'après Funke et coll. (2004), il existe une association allélique significative pour deux des SNP à l'étude. Dans leur étude, l'allèle T du SNP rs1018381 et l'allèle A du SNP rs2619528 se sont montrés plus fréquents chez les cas que chez les témoins. Les rapports de cotes ont été respectivement estimés à 1.76 et à 1.41 ; de plus, les seuils observés des tests ont été de 0.0026 et 0.0171.

Par ailleurs, Funke et coll. (2004) observent que l'haplotype $C\mathbf{T}CT\mathbf{A}C$ est significativement plus fréquent chez les cas que chez les témoins, au seuil de 0.005.

5.2 Résultats d'une étude maison

L'étude d'association réalisée dans le cadre de ce mémoire a été effectuée à partir d'un échantillon de cas-témoins de la population québécoise. Plusieurs SNP dans les régions des gènes NRG1 et DTNBP1 ont été génotypés. Pour le gène NRG1, on possède l'information génétique des cinq mêmes SNP étudiés par Stefansson et coll. (2003). Pour le gène DTNBP1, on dispose de l'information génétique de trois des six SNP étudiés par Funke et coll. (2004), soit rs1018381, rs2619528 et rs1011313. L'échantillon compte 313 cas et 253 témoins. Les associations allélique et haplotypique des SNP ont été testées.

5.2.1 Résultats des tests alléliques

La présence d'association allélique a été vérifiée à partir de tests du khi-deux. Pour ce faire, la procédure `casecontrol` du module `genetics` du logiciel SAS a été utilisée. Les résultats des tests d'association sur les allèles et sur les génotypes sont respectivement présentés dans les tableaux 5.2 et 5.3. Noter que les tests sur les génotypes n'ont pas été présentés dans les deux études décrites à la section 5.1.

gène	SNP	allèle[a]	fréquences cas ($N = 313$)[b]	fréquences témoins ($N = 253$)[b]	rapport de cotes[c]	test du χ^2 seuil observé
NRG1	nrg221132	*A*	0.1283	0.1245	0.97	0.8494
		G	0.8717	0.8755		
	nrg221533	*T*	0.6517	0.6384	0.94	0.6517
		C	0.3483	0.3616		
	nrg241930	**G**	0.6382	0.6455	0.97	0.8047
		T	0.3618	0.3545		
	nrg243177	**T**	0.3703	0.3971	0.89	0.3685
		C	0.6297	0.6029		
	nrg433E1006	**G**	0.8983	0.8951	1.04	0.8636
		A	0.1017	0.1049		
DTNBP1	rs1018381	*C*	0.9281	0.9265	0.98	0.9203
		T	0.0719	0.0735		
	rs2619528	**A**	0.1759	0.1660	1.07	0.6646
		G	0.8241	0.8340		
	rs1011313	*C*	0.9003	0.8996	0.99	0.9677
		T	0.0997	0.1004		

[a] En gras, l'allèle plus fréquent chez les cas que chez les témoins dans les études antérieures.
[b] Pour certains SNP, les valeurs de N peuvent varier. Au plus, 30 sujets sont perdus, soit environ 5% de l'échantillon total.
[c] Les rapports de cotes indiquent l'association des allèles plus fréquents chez les cas que chez les témoins dans les études antérieures.

TAB. 5.2 – Résultats des tests d'association sur les allèles pour l'échantillon de cas-témoins du CRULRG

Les fréquences alléliques des cas et des témoins de l'échantillon du CRULRG ne sont pas significativement différentes. Par ailleurs, les allèles associés dans les études de Stefansson et coll. (2003) et de Funke et coll. (2004) sont ici un peu moins fréquents chez les cas que chez les témoins, sauf pour le SNP rs2619528. On ne détecte donc aucune association allélique.

Les fréquences génotypiques des cas et des témoins ne sont pas significativement différentes. Le SNP rs2619528 du gène DTNBP1 semble le plus susceptible d'être associé à la SZ, le seuil observé du test étant de 0.2288. On remarque que la fréquence du génotype A/A est deux fois plus grande chez les cas que chez les témoins ($0.0456 > 0.0205$). Ceci concorde avec les résultats de Funke et coll. (2004), lesquels trouvaient un rapport de cotes de 1.41 pour l'allèle A du même SNP.

Le tableau 5.4 présente une autre version du test d'association basée sur les gé-

gène	SNP	génotype[a]	fréquences cas ($N = 313$)[b]	fréquences témoins ($N = 253$)[b]	rapport de cotes[c]	test du χ^2 seuil observé
NRG1	nrg221132	A/A	0.0133	0.0204	1.14	0.6910
		A/G	0.2300	0.2082		
		G/G	0.7567	0.7714		
	nrg221533	**T/T**	0.4276	0.4050	1.10	0.8698
		T/C	0.4483	0.4669		
		C/C	0.1241	0.1281		
	nrg241930	G/G	0.3993	0.4057	1.08	0.9570
		G/T	0.4778	0.4795		
		T/T	0.1229	0.1148		
	nrg243177	T/T	0.1331	0.1440	1.20	0.5951
		T/C	0.4744	0.5062		
		C/C	0.3925	0.3498		
	nrg433E1006	**G/G**	0.8034	0.7984	1.03	0.9776
		G/A	0.1897	0.1934		
		A/A	0.0069	0.0082		
DTNBP1	rs1018381	C/C	0.8595	0.8612	1.06	0.7247
		C/T	0.1373	0.1306		
		T/T	0.0033	0.0082		
	rs2619528	**A/A**	0.0456	0.0205	2.28	0.2288
		A/G	0.2606	0.2910		
		G/G	0.6938	0.6885		
	rs1011313	C/C	0.8072	0.8156	1.13	0.4784
		C/T	0.1863	0.1680		
		T/T	0.0065	0.0164		

[a] En gras, le génotype plus fréquent chez les cas que chez les témoins dans l'étude du CRULRG.
[b] Pour certains SNP, les N peuvent varier. Au plus, 30 sujets sont perdus, soit environ 5% de l'échantillon total.
[c] Les rapports de cotes indiquent l'association des génotypes plus fréquents chez les cas que chez les témoins dans l'étude du CRULRG.

TAB. 5.3 – Résultats des tests d'association sur les génotypes pour l'échantillon de cas-témoins du CRULRG

notypes. Celle-ci permet de comparer plus spécifiquement les fréquences du génotype A/A des cas et des témoins. Le principe se résume à regrouper dans la même catégorie les fréquences des génotypes A/G et G/G. On ne détecte toujours pas de différence significative entre les fréquences génotypiques des cas et des témoins, au seuil de 5%. Cependant, au vu des résultats de l'étude de Funke et coll. (2004), on pourrait choisir la contre-hypothèse unilatérale suivante :

H_1 : le génotype A/A est plus fréquent chez les cas que chez les témoins.

SNP	génotype	fréquences cas	fréquences témoins	test du χ^2 seuil observé	rapport de cotes (I.C.)
rs2619528	A/A	0.0456	0.0205	0.1086	2.2840
	A/G et G/G	0.9544	0.9795		(0.8111–6.4316)

TAB. 5.4 – Résultat du test d'association pour le génotype A/A du SNP rs2619528

Le seuil observé obtenu serait alors réduit de moitié et on aurait une valeur de 0.0543. Ce résultat suggère la présence d'une association entre le SNP rs2619528 et la SZ, tel que rapporté par Funke et coll. (2004).

5.2.2 Résultats des tests haplotypiques

La présence d'association haplotypique a été éprouvée à partir de tests du rapport des vraisemblances (TRV). La procédure `haplotype` du module `genetics` du logiciel SAS a été utilisée à cette fin. Les fréquences hapotypiques ont été estimées par l'algorithme EM.

Le nombre d'haplotypes possibles pour les cinq SNP du gène NRG1 est $2^5 = 32$. Pour les trois SNP du gène DTNBP1, on en compte $2^3 = 8$. Quand le nombre d'haplotypes est grand, il arrive que certains d'entre eux soient très rares ou qu'ils ne soient pas observés dans l'échantillon recruté. Le tableau 5.5, qui rapporte les résultats des tests d'association haplotypique, ne présente que les fréquences des haplotypes observés chez au moins 1% des sujets de l'échantillon total.

Comme l'illustre le tableau 5.5, les fréquences haplotypiques des cas et des témoins de l'échantillon du CRULRG sont pratiquement égales. Les seuils observés des tests globaux sont très grands. On n'a donc aucune raison de croire que les haplotypes des SNP étudiés sont associés à la SZ. Les tests spécifiques ont été présentés mais, étant donné que les tests globaux ne permettent pas de rejeter l'hypothèse nulle de non association, on ne devrait pas les interpréter.

Notons que les haplotypes $GCGTG$ et TAC, dont on a vu qu'ils sont associés dans les deux études présentées à la section 5.1, sont dans ce cas-ci un peu moins fréquents chez les cas que chez les témoins, ce qui semble même contredire les études antérieures. On ne reproduit donc pas les résultats de ces études. Différentes explications sont pos-

gène	haplotype[a]	fréquences		TRV global	TRV spécifique
		cas	témoins	seuil	seuil
		($N = 313$)	($N = 253$)	observé	observé
NRG1	*ATTCG*	0.1130	0.1123	0.9962	0.7843
	GTGTG	0.0425	0.0339		0.4621
	GTGTA	0.0129	0.0195		0.3850
	GTGCG	0.2443	0.2268		0.4807
	GTGCA	0.0151	0.0181		0.6457
	GTTCG	0.1528	0.1541		0.8994
	GTTCA	0.0593	0.0633		0.7830
	GCGTG	0.2941	0.3238		0.2901
	GCGCG	0.0149	0.0161		0.6978
	GCTCG	0.0246	0.0086		0.0482
DTNBP1	*CAC*	0.1036	0.0914	0.9998	0.4982
	CGC	0.7248	0.7347		0.6995
	CGT	0.0997	0.0996		0.9673
	TAC	0.0723	0.0735		0.9425

[a] Seuls les haplotypes ayant une fréquence > 0.01 dans l'échantillon combiné apparaissent dans le tableau.

TAB. 5.5 – Résultats des tests d'association haplotypique pour l'échantillon de cas-témoins du CRULRG

sibles. Les deux principales raisons pourraient être que la définition de la maladie n'est pas exactement la même dans les différentes études ou que plusieurs gènes soient responsables de la SZ et que l'importance de chacun d'eux varie d'une population à l'autre.

Chapitre 6

Description de tests d'association dans un échantillon de trios

Le principe général des études d'association dans un échantillon de trios se résume à considérer les allèles transmis par les parents à l'enfant malade comme étant des allèles *cas* et les allèles non transmis comme étant des allèles *témoins*. Ce type de devis permet d'éviter les problèmes de stratification de la population puisque les allèles cas et témoins sont parfaitement appariés. Pour illustrer les tests présentés dans ce chapitre, l'échantillon de cas-témoins du tableau 4.1 a été converti en échantillon de trios. Pour ce faire, on utilise les génotypes des cas comme étant ceux des enfants malades des trios. Obligatoirement, cela identifie du même coup les allèles transmis par chacun des deux parents d'un trio. Les génotypes des témoins ont ensuite servi à compléter les génotypes des parents tel qu'illustré à la figure 6.1 et, dans les trios, ils représentent donc les allèles non transmis.

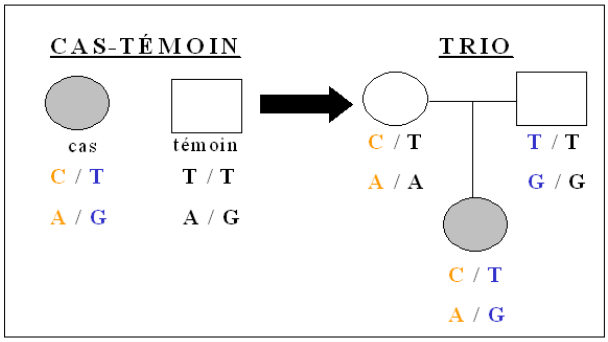


FIG. 6.1 – *Illustration de la conversion d'un échantillon de cas-témoins en échantillon de trios*

Cette procédure peut conduire à des échantillons de trios différents. Les données utilisées pour les exemples proviennent de l'échantillon de trios présenté au tableau 6.1. Cet échantillon a été obtenu en appariant aléatoirement les cas et les témoins du tableau 4.1.

	père		mère		enfant	
trio	SNP 1	SNP 2	SNP 1	SNP 2	SNP 1	SNP 2
1	C/T	A/A	C/C	A/G	C/C	A/A
2	C/C	A/A	C/C	A/A	C/C	A/A
3	C/T	A/G	C/T	A/A	C/C	A/A
4	C/T	A/G	C/T	A/G	C/C	A/A
5	C/T	A/G	C/C	A/A	C/T	A/G
6	C/C	A/A	C/C	A/A	C/C	A/A
7	C/C	A/A	C/C	A/A	C/C	A/A
8	C/C	A/A	C/C	A/A	C/C	A/A
9	C/T	A/G	C/C	A/A	C/C	A/A
10	C/C	A/G	C/C	A/A	C/C	A/A
11	C/T	A/G	C/C	A/G	C/C	A/A
12	C/T	A/G	C/T	A/G	C/C	A/A
13	C/C	A/A	C/C	A/A	C/C	A/A
14	C/C	A/A	C/T	A/G	C/C	A/A
15	C/T	A/G	C/T	A/G	C/C	A/A

TAB. 6.1 – Génotypes des 15 trios pour les SNP 1 et 2 formés à partir des allèles observés chez les cas et témoins du tableau 4.1

6.1 Test d'association allélique

6.1.1 Le test de déséquilibre de transmission

En présence d'un échantillon de trios, le test de déséquilibre de transmission (TDT) est très souvent utilisé pour détecter la présence d'association allélique. Le TDT a été introduit par Spielman et coll. (1993). Ce test s'appuie sur l'observation des allèles transmis par les parents à l'enfant atteint. On teste l'hypothèse

H_0 : les allèles du marqueur ne sont pas associés à la maladie.

Cette hypothèse peut être éprouvée en comparant la distribution des allèles transmis par les parents à celle des allèles non transmis. Sous H_0, on s'attend à ce que ces distributions coïncident.

Soient n trios, pour lesquels on compte n enfants et $2n$ parents. Supposons que l'on étudie l'association d'un marqueur à deux allèles, $A1$ et $A2$. Le TDT permet de tenir compte de l'appariement entre les allèles transmis et les allèles non transmis. On crée un tableau de fréquences, tel qu'illustré au tableau 6.2.

		allèles non transmis		
		$A1$	$A2$	
allèles transmis	$A1$	a	b	$a+b$
	$A2$	c	d	$c+d$
		$a+c$	$b+d$	$2n$

TAB. 6.2 – Combinaison des allèles $A1$ et $A2$ transmis et non transmis par les $2n$ parents des n enfants malades

Les valeurs du tableau 6.2 s'interprètent comme suit :

a = nombre de parents ayant le génotype $A1/A1$ qui ont transmis l'allèle $A1$;

b = nombre de parents ayant le génotype $A1/A2$ qui ont transmis l'allèle $A1$;

c = nombre de parents ayant le génotype $A1/A2$ qui ont transmis l'allèle $A2$;

d = nombre de parents ayant le génotype $A2/A2$ qui ont transmis l'allèle $A2$.

On remarque que seuls les parents ayant des génotypes hétérozygotes ($A1/A2$) sont informatifs.

Le TDT est basé sur le test de McNemar (McNemar, 1947). Il s'agit d'une procédure conditionnelle, dans laquelle le nombre $m = b + c$ de paires discordantes est considéré comme fixé. Sous H_0, la variable aléatoire b suit alors une loi binomiale :

$$b \sim \mathcal{BIN}(m, 1/2).$$

On s'attend alors à ce que b soit relativement proche de $m/2$, de sorte qu'alors $b \approx c$. Si la valeur $m/4$ est supérieure ou égale à 5, la distribution de b peut être approximée par une loi normale, à savoir :

$$b \approx \mathcal{N}\left(\frac{m}{2}, \frac{m}{4}\right).$$

Ceci nous amène à une statistique de la forme suivante :

$$Z = \frac{1}{\sqrt{m/4}} \left(b - \frac{m}{2} \right)$$

En élevant Z au carré, on obtient la statistique

$$T = \frac{(b-c)^2}{b+c},$$

qui, sous H_0, suit approximativement une distribution du khi-deux à un degré de liberté.

Un TDT peut être effectué par l'entremise de la procédure `family` de SAS.

6.1.2 Exemple de test de déséquilibre de transmission

Testons l'association des allèles du SNP 1 avec la maladie pour l'échantillon de trios du tableau 6.1. On peut utiliser pour ce faire les données du tableau 6.3.

		allèles non transmis		
		C	T	
allèles transmis	C	17	12	29
	T	1	0	1
		18	12	30

TAB. 6.3 – Combinaison des allèles C et T transmis et non transmis par les 30 parents des 15 enfants malades

La statistique du test est la suivante :

$$T = \frac{(12-1)^2}{12+1} = \frac{121}{13} = 9.31.$$

Le seuil asymptotique associé à cette statistique vaut 0.0023, ce qui conduit au rejet de l'hypothèse nulle. On en conclut que les parents ayant un génotype hétérozygote (C/T) transmettent significativement plus souvent l'allèle C aux enfants. En d'autres termes, l'allèle C est associé à la maladie.

6.2 Test d'association haplotypique

Récemment, des chercheurs du Département de biostatistique de l'École de santé publique de Harvard ont développé une méthodologie générale pour tester l'association à

partir d'échantillons de familles (Laird et coll., 2000). La classe de tests qu'ils proposent est couramment appelée *Family-based association tests* (FBAT).

Les tests FBAT peuvent être effectués à partir d'un logiciel du même nom que l'on peut télécharger gratuitement (FBAT, 2006). Ces tests permettent entre autres de détecter l'association entre une maladie et des haplotypes. Afin de faciliter la compréhension de ce volet qui nous intéresse plus particulièrement, la théorie générale des tests FBAT sera d'abord présentée.

6.2.1 Théorie générale de FBAT

6.2.1.1 Contexte et hypothèse du test

Considérons n familles nucléaires, c'est-à-dire composées chacune des parents biologiques et des enfants. On suppose ces n familles nucléaires indépendantes. Soit n_i, le nombre d'enfants dans la famille i. On teste l'hypothèse

H_0 : il n'y a ni liaison ni association entre les marqueurs et la maladie.

La statistique du test pour la famille i est définie par l'équation suivante :

$$S_i = \sum_{j=1}^{n_i} X_{ij} T_{ij},$$

où

X_{ij} = une certaine fonction du génotype de l'enfant j de la famille i ;
T_{ij} = une certaine fonction du trait de l'enfant j de la famille i.

6.2.1.2 Précisions sur les fonctions X et T

La fonction X du génotype peut être sous forme de scalaire ou de vecteur. Par exemple, pour un marqueur à trois allèles, disons A, B et C, on peut choisir de compter le nombre d'allèles A que possède chaque individu. Dans ce cas, X est sous forme de scalaire et se définit ainsi :

$$X = \begin{cases} 2, & \text{si le génotype de l'individu est } A/A\,; \\ 1, & \text{si le génotype de l'individu est } A/B \text{ ou } A/C\,; \\ 0, & \text{si le génotype de l'individu est } B/B,\ B/C \text{ ou } C/C. \end{cases}$$

Pour le même marqueur, on peut opter pour une codification de X sous forme de vecteur tel que présenté dans le tableau 6.4.

	vecteur X	
génotype de l'individu	génotype	allèle
A/A	(0,0,0,0,0)	(2,0,0)
A/B	(1,0,0,0,0)	(1,1,0)
A/C	(0,1,0,0,0)	(1,0,1)
B/B	(0,0,1,0,0)	(0,2,0)
B/C	(0,0,0,1,0)	(0,1,1)
C/C	(0,0,0,0,1)	(0,0,2)

TAB. 6.4 – Exemples de codes numériques pour un X sous forme de vecteur (Schaid, 1996)

La fonction T du trait s'obtient à partir de l'équation $T_{ij} = Y_{ij} - \mu$. Ici, Y_{ij} correspond au trait, qui peut être dichotomique (0 : sain, 1 : malade) ou continu (par exemple, l'indice de masse corporelle), et μ est un paramètre d'ajustement (par exemple, la moyenne de Y). Mentionnons également que la valeur T_{ij} peut être ajustée pour des covariables. Si l'on introduit la covariable Z, l'équation pourrait être de la forme suivante :

$$T_{ij} = Y_{ij} - \beta_0 - \beta_1 Z_{ij}.$$

6.2.1.3 Statistique du test

On définit une fonction score

$$U = \sum_{i=1}^{n} \{S_i - \mathrm{E}(S_i)\}$$

et une variance

$$V = \sum_{i=1}^{n} \mathrm{var}(S_i).$$

La forme du test dépend de la codification choisie pour X. Si X est un scalaire, le test est de la forme suivante :

$$Z = \frac{U}{\sqrt{V}}\ .$$

Sous H_0, la statistique Z obéit approximativement à une loi normale centrée réduite. Mentionnons qu'en élevant Z au carré, sa distribution peut alors être approximée par une loi du khi-deux à un degré de liberté.

Si X est un vecteur, on opte plutôt pour un test de la forme qui suit :

$$X^2 = U^\top V^{-1} U.$$

Sous H_0, la statistique X^2 obéit approximativement à une loi du khi-deux à ν degrés de liberté, où ν représente le rang de V.

6.2.1.4 Calcul de l'espérance et de la variance de S_i

Tel que décrit dans le rapport technique de Horvath et coll. (2000), le calcul de l'espérance et de la variance de S_i se fait en conditionnant sur les génotypes qui peuvent être observés chez les enfants de la famille i. On obtient l'espérance à partir de l'équation suivante :

$$\mathrm{E}(S_i) = \sum_{j=1}^{n_i} \mathrm{E}\left(X_{ij}\right) T_{ij},$$

où

$$\mathrm{E}(X_{ij}) = \sum_{g \in G} X(g)\mathrm{P}(g).$$

Ici, G correspond à l'ensemble de tous les génotypes possibles pour les enfants de la famille i, tandis que $X(g)$ représente la fonction du génotype g et $\mathrm{P}(g)$ réfère à la probabilité que l'enfant ait le génotype g sachant qu'il fait partie de la famille i.

Pour la variance, on procède comme suit :

$$\mathrm{var}(S_i) = \sum_j T_{ij}^2 \mathrm{var}(X_{ij}) + \sum_j \sum_{k \neq j} T_{ij} T_{ik} \mathrm{cov}(X_{ij}, X_{ik}),$$

où

$$\begin{aligned} \mathrm{var}(X_{ij}) &= \mathrm{E}(X_{ij}^2) - \mathrm{E}(X_{ij})^2 \\ &= \sum_{g \in G} X(g)X(g)^\top \mathrm{P}(g) - \sum_{g \in G} X(g)\mathrm{P}(g) \sum_{\tilde{g} \in G} X(\tilde{g})^\top \mathrm{P}(\tilde{g}) \end{aligned}$$

et

$$\begin{aligned}\mathrm{cov}(X_{ij}, X_{ik}) &= \mathrm{E}(X_{ij}X_{ik}) - \mathrm{E}(X_{ij})\mathrm{E}(X_{ik}) \\ &= \sum_{g\in G}\sum_{\tilde{g}\in G} X(g)X(\tilde{g})^\top \mathrm{P}(g\tilde{g}) - \sum_{g\in G} X(g)\mathrm{P}(g)\sum_{\tilde{g}\in G} X(\tilde{g})^\top \mathrm{P}(\tilde{g}).\end{aligned}$$

6.2.1.5 Remarques

Le TDT est un cas particulier des tests FBAT. Si on a un seul enfant par famille (un enfant atteint de la maladie), un trait dichotomique (0 : sain, 1 : malade), un paramètre d'ajustement nul, un seul marqueur et un X qui compte le nombre d'un certain type d'allèle à ce marqueur, on retrouve exactement la statistique du TDT :

$$(Z_{FBAT})^2 = T_{TDT}.$$

Il est possible de tester une autre hypothèse nulle, à savoir l'hypothèse H_0 : il y a liaison entre les marqueurs et la maladie, mais pas association. Cette approche peut être utile quand on a identifié par étude de liaison des marqueurs liés à la maladie et que l'on veut déterminer s'il y a également présence d'association. Le test diffère de celui pour l'hypothèse nulle de non liaison et non association, car la variance doit tenir compte de la dépendance entre les familles nucléaires appartenant au même pedigree[1]. En effet, la présence de la liaison entraîne de la corrélation entre les génotypes des sujets provenant d'un même pedigree.

6.2.2 Exemple de test d'association haplotypique avec FBAT

Tout d'abord, il est important de mentionner que les tests d'association haplotypique avec FBAT font l'hypothèse qu'il n'y a pas de recombinaison possible entre les marqueurs qui forment les haplotypes à l'étude. Cela implique qu'un parent transmettra un haplotype identique à celui reçu par un de ses propres parents.

Intéressons-nous à l'association des haplotypes dans les familles de l'exemple du tableau 6.1. Rappelons le contexte de cet exemple. On dispose de $n = 15$ familles qui sont toutes des trios avec un seul enfant par famille, un enfant malade, et pour lesquelles on a l'information génétique de deux SNP à deux allèles chacun. Les haplotypes sont : CA, CG, TA et TG. Le trait étudié est la présence de la maladie.

[1] Le terme pedigree est utilisé pour définir une famille dans un sens plus large, c'est-à-dire constituée de plusieurs familles nucléaires reliées.

Supposons que l'on choisit un paramètre d'ajustement égal à 0 et que l'on codifie X sous forme de vecteur qui compte le nombre d'haplotypes CA, CG, TA et TG. La statistique du test se simplifie alors considérablement. Puisqu'on a un seul enfant par famille, l'indice j et la covariance entre les enfants d'une même famille disparaissent. De plus, la fonction du trait T étant la même pour tous les trios, on trouve $T_i = Y_i = 1$ pour tout i.

Dans le cadre de test d'association haplotypique, on calcule l'espérance et la variance de S_i sachant les génotypes observés chez les individus de la famille i. Par conséquent, les fonctions U et V de la statistique du test sont conditionnelles aux génotypes des individus de l'échantillon. L'espérance et la variance de S_i s'obtiennent à partir des équations suivantes :

$$\begin{aligned} \mathrm{E}(S_i) &= \mathrm{E}(X_i) = \sum_{g \in G} X(g)\mathrm{P}(g); \\ \mathrm{var}(S_i) &= \mathrm{var}(X_i) = \sum_{g \in G} X(g)X(g)^\top \mathrm{P}(g) - \sum_{g \in G} X(g)\mathrm{P}(g) \sum_{\tilde{g} \in G} X(\tilde{g})^\top \mathrm{P}(\tilde{g}). \end{aligned}$$

Précisons que G correspond ici à l'ensemble de toutes les paires d'haplotypes possibles pour les enfants de la famille i, $X(g)$ représente la fonction de la paire d'haplotypes g et $\mathrm{P}(g)$ réfère à la probabilité que l'enfant ait la paire d'haplotypes g sachant qu'il fait partie de la famille i.

Pour cet exemple particulier, on connaît les haplotypes de tous les sujets de l'échantillon. Pour les sujets homozygotes à l'un des deux SNP ou aux deux SNP, on retrouve directement les haplotypes. Pour les sujets hétérozygotes aux deux SNP, on arrive à les déduire à partir des génotypes des parents ou de l'enfant puisqu'on suppose qu'il n'y a pas de recombinaison possible.

Pour tous les enfants, sauf celui du trio 5, on a $X_i^\top = (2, 0, 0, 0)$. Tel qu'illustré à la figure 6.2, l'enfant du trio 5 a les haplotypes CA et TG. Par conséquent, son vecteur X est : $X_5^\top = (1, 0, 0, 1)$.

Ainsi, en général, les échantillons de familles fournissent beaucoup plus d'information permettant la reconstruction des haplotypes que les échantillons de cas-témoins et ces haplotypes peuvent souvent être déterminés sans ambiguïté. Cependant, il demeure possible que les haplotypes soient difficilement identifiables dans les familles, par exemple en présence de données manquantes. Il existe dans ce cas une version de l'algorithme EM pour les échantillons de familles.

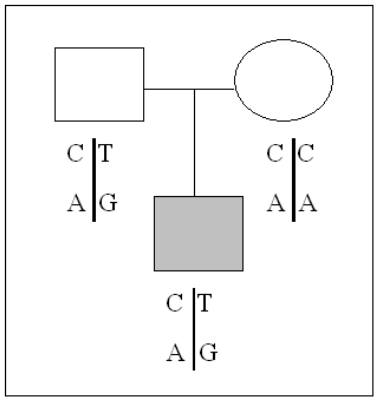


FIG. 6.2 – *Haplotypes des individus du trio 5*

6.2.2.1 Détails du calcul de l'espérance et de la variance de S_i

Illustrons les détails du calcul de $\mathrm{E}(S_i)$ et de $\mathrm{var}(S_i)$ pour le trio 5. Le père peut transmettre à l'enfant l'haplotype CA ou TG, avec une probabilité de 1/2 pour chaque haplotype. La mère ne peut transmettre qu'un haplotype, l'haplotype CA, et ce avec une probabilité de 1. Le tableau qui suit présente les deux paires d'haplotypes possibles pour l'enfant du trio 5 avec les probabilités associées à chacun :

g	CA, CA	CA, TG
$X(g)$	(2,0,0,0)	(1,0,0,1)
$P(g)$	0.5	0.5

Il s'ensuit que

$$\mathrm{E}(S_5) = \mathrm{E}(X_5) \;=\; \sum_{g \in G} X(g)\mathrm{P}(g) = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} \frac{1}{2} + \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \frac{1}{2} = \begin{pmatrix} 1.5 \\ 0 \\ 0 \\ 0.5 \end{pmatrix}.$$

De même,

$$\begin{aligned}\mathrm{var}(S_5) = \mathrm{var}(X_5) &= \sum_{g\in G} X(g)X(g)^\top \mathrm{P}(g) - \sum_{g\in G} X(g)\mathrm{P}(g) \sum_{\tilde{g}\in G} X(\tilde{g})^\top \mathrm{P}(\tilde{g}) \\ &= \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} (2,0,0,0)\,\frac{1}{2} + \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} (1,0,0,1)\,\frac{1}{2} \\ &\quad - \begin{pmatrix} 1.5 \\ 0 \\ 0 \\ 0.5 \end{pmatrix} (1.5,0,0,0.5)\end{aligned}$$

et donc

$$\mathrm{var}(S_5) = \quad = \begin{pmatrix} 0.25 & 0 & 0 & -0.25 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -0.25 & 0 & 0 & 0.25 \end{pmatrix}.$$

6.2.2.2 Résultats du test

Test global

On teste l'hypothèse

H_0 : aucun haplotype n'est associé à la maladie.

Puisque X est un vecteur, le test est de la forme $X^2 = U^\top V^{-1} U$, où

$$U = \begin{pmatrix} 7 \\ -1.5 \\ -1 \\ -4.5 \end{pmatrix} \quad \text{et} \quad V = \begin{pmatrix} 6.5 & -1 & -0.75 & -4.75 \\ -1 & 0.75 & 0 & 0.25 \\ -0.75 & 0 & 0.5 & 0.25 \\ -4.75 & 0.25 & 0.25 & 4.25 \end{pmatrix}.$$

L'inverse matriciel de V est

$$V^{-1} = \begin{pmatrix} 0.2088 & -0.1211 & -0.2861 & 0.1985 \\ -0.1211 & 0.9098 & -0.5747 & -0.2139 \\ -0.2861 & -0.5747 & 1.2809 & -0.4201 \\ 0.1985 & -0.2139 & -0.4201 & 0.4356 \end{pmatrix}.$$

La statistique du test global est $X^2 = 8.0309$ et elle peut être comparée, sous l'hypothèse nulle, à une loi du khi-deux ayant trois degrés de liberté. Le seuil observé vaut 0.0454. Au seuil de 5%, on conclut donc à la présence d'une association entre la maladie et au moins un des haplotypes.

Tests spécifiques

On teste cette fois une hypothèse plus spécifique, soit

$$H_0 : \text{l'haplotype } j \text{ n'est pas associé à la maladie.}$$

Les tests à un haplotype particulier de FBAT sont de la même forme que ceux pour le test du score, à savoir :

$$Z_j = \frac{U[j]}{\sqrt{V[j,j]}} .$$

Sous H_0, Z_j obéit approximativement à une loi normale centrée réduite.

Pour l'exemple, on obtient les résultats suivants :

$$\begin{aligned} Z_{CA} &= \frac{U[1]}{\sqrt{V[1,1]}} = \frac{7}{\sqrt{6.5}} = 2.746, \\ Z_{CG} &= \frac{U[2]}{\sqrt{V[2,2]}} = \frac{-1.5}{\sqrt{0.75}} = -1.732, \\ Z_{TA} &= \frac{U[3]}{\sqrt{V[3,3]}} = \frac{-1}{\sqrt{0.5}} = -1.414, \\ Z_{TG} &= \frac{U[4]}{\sqrt{V[4,4]}} = \frac{-4.5}{\sqrt{4.25}} = -2.183. \end{aligned}$$

Les seuils observés sont respectivement 0.0060, 0.0833, 0.1573 et 0.0290. Au seuil de 5%, on conclut donc que l'haplotype CA est positivement associé à la maladie alors que l'haplotype TG l'est négativement. Si l'on choisit d'utiliser une correction de Bonferroni, seul le résultat pour l'haplotype CA demeure, puisque le seuil observé pour l'haplotype TG est supérieur à 1.25%.

Maintenant que l'on est plus familier avec les différentes stratégies d'analyse qui existent pour l'étude d'association génétique, on s'intéressera au prochain chapitre à l'efficacité de ces méthodes.

Chapitre 7

Efficacité des tests d'association en génétique

On a vu aux chapitres précédents qu'il existe plusieurs méthodes de détection d'une association génétique. Il a été question de deux devis d'échantillonnage possibles, l'échantillonnage de cas-témoins et l'échantillonnage de trios. On a aussi vu que l'association génétique peut être testée au niveau allélique et au niveau haplotypique.

Existe-il une stratégie d'analyse plus efficace que les autres pour détecter l'association génétique ? Pour tenter de répondre à cette question, on s'intéressera d'abord à la comparaison des deux devis d'échantillonnage, cas-témoins versus trios, puis à la comparaison des deux approches analytiques, allélique versus haplotypique.

Il convient dans un premier temps de déterminer les avantages et inconvénients de chaque stratégie d'analyse, ce qui fait l'objet de la section 7.1. La comparaison de la puissance statistique des différentes stratégies s'avère également une bonne approche pour investiguer le sujet. Une étude de simulation a d'ailleurs été réalisée à cet effet. Les détails en sont présentés au chapitre 8.

Dernièrement, divers scientifiques se sont penchés sur la question de l'efficacité des différents tests d'association génétique. À la section 7.2, on résume quelques-uns de leurs travaux, identifiés après revue de littérature.

7.1 Avantages et inconvénients des différentes stratégies d'analyse

Le tableau 7.1 résume les principales caractéristiques des devis de cas-témoins et de trios. Le tableau 7.2 présente celles des approches analytiques allélique et haplotypique.

devis	avantages	inconvénients
cas-témoins	- facile à recruter ; - peu coûteux.	- sélection soignée des témoins ; - problème de stratification de population (voir l'exemple à la section 4.3.1).
trios	- un échantillon utilisé pour des études de liaison peut être repris ; - pas besoin de recruter des témoins ; - robuste à la stratification de population.	- plus coûteux que le devis de cas-témoins (1.5 fois plus de gens à génotyper) ; - peut être difficile à recueillir ; - peut y avoir des données manquantes (famille dispersée, un parent décédé ou difficile à rejoindre).

Tab. 7.1 – Avantages et inconvénients des devis d'échantillonnage

approche	avantages	inconvénients
allélique	- analyses simples (tests du khi-deux). - ne nécessite pas l'estimation des fréquences haplotypiques.	- ne permet pas de tenir compte du DL entre les marqueurs ; - problème de tests multiples.
haplotypique	- permet de diminuer le nombre de tests ; - permet de capturer l'information sur le DL entre les marqueurs.	- les degrés de liberté augmentent avec le nombre d'haplotypes ; - on doit estimer les fréquences haplotypiques.

Tab. 7.2 – Avantages et inconvénients des approches analytiques

7.2 Revue de littérature

Une revue de littérature a permis de repérer quelques articles récents qui concernent la comparaison des différentes stratégies d'analyse pour l'étude d'association génétique.

7.2.1 Étude de Hintsanen et al. (2006)

Récemment, Hintsanen et coll. (2006) ont publié une étude de simulation visant à comparer l'efficacité des devis de trios et de cas-témoins pour détecter l'association haplotypique. La comparaison des approches allélique et haplotypique n'a pas été considérée dans leur étude.

Sous certaines conditions de recombinaison et de mutation, ils ont simulé la transmission d'information génétique dans des familles comptant 15 générations. Leur population totale comprenait 100 000 individus. Pour les tests d'association, ils ont prélevé des échantillons de 500 sujets indépendants (250 cas et 250 témoins), ainsi que des échantillons de 167 trios. Ainsi, le nombre de sujets génotypés est le même pour les deux types d'échantillon, mais les échantillons de trios comportent moins de cas. Le nombre de marqueurs étudiés variait de 1 à 10.

Dans le cadre du présent mémoire, on a plutôt opté pour une approche simple avec une simulation plus directe et un nombre de marqueurs plus petit. Nos échantillons sont également de plus petite taille et on a choisi un même nombre de cas dans les deux types d'échantillon plutôt qu'un même nombre de sujets génotypés.

Pour étudier l'association des haplotypes, Hintsanen et coll. (2006) ont effectué de simples tests du khi-deux, ce qui est peu commun pour ce type d'analyse. Dans le présent travail, on se propose plutôt d'étudier l'association à partir de tests plus couramment utilisés, soit le test du rapport des vraisemblances, le test du score et les tests FBAT.

Pour estimer les fréquences alléliques dans les échantillons de cas-témoins, l'approche `HaploRec` à base de chaînes de Markov a été privilégiée par Hintsanen et coll. (2006). De notre côté, on a préféré l'algorithme EM. Dans les échantillons de trios, ces auteurs n'ont pas inféré les haplotypes. Si l'identification des haplotypes ne se faisait pas avec certitude pour un sujet, une donnée manquante lui était attribuée. Dans notre étude, une adaptation de l'algorithme EM, implantée dans les tests FBAT, permet de contourner cette difficulté sans perte d'information.

La conclusion générale de Hintsanen et coll. (2006) est que le devis de cas-témoins est plus puissant que le devis de trios. Ils expliquent cette différence en partie par le plus petit nombre de cas dans les échantillons de trios, 1/3 de moins, et la perte d'information dans les trios engendrée par l'attribution de données manquantes.

7.2.2 Études réalisées dans le cadre du GAW n° 14

Le « Genetic Analysis Workshop »(GAW) est un atelier qui a pour but d'évaluer et de comparer des méthodes d'analyse statistique en génétique. Des participants au GAW n° 14 ont étudié l'efficacité des tests d'association (Bull et coll., 2005).

D'une part, Pinnaduwage et Briollais (2005) ont comparé des tests d'association allélique et haplotypique. Ils disposaient d'un échantillon de 143 familles provenant d'une étude sur la génétique de l'alcoolisme. Pour tester l'association allélique, ils ont choisi le TDT et pour tester l'association haplotypique, ils ont opté pour les tests FBAT. Les deux stratégies d'analyse ont été effectuées par l'entremise du logiciel FBAT. Les auteurs ont considéré trois marqueurs à la fois pour les haplotypes. Les fréquences haplotypiques ont été estimées à partir de l'algorithme implanté dans le logiciel FBAT. Des tests globaux et spécifiques ont été faits pour tester l'association des haplotypes.

D'autre part, Shephard et coll. (2005) ont travaillé à partir de données simulées. Ils ont comparé à la fois les deux approches analytiques et les deux devis d'échantillonnage. Chaque échantillon de cas-témoins comportait 100 cas et 50 témoins et chaque échantillon de trios était composé de 100 trios. Pour détecter la présence d'association allélique, le test du khi-deux a été utilisé dans les cas-témoins et le TDT dans les trios. Pour tester l'association haplotypique, la procédure « haplotype trend regression »(Zaykin et coll., 2002) a été utilisée dans les cas-témoins et une version du TDT pour les haplotypes (Dudbridge et coll., 2000) dans les trios. Les auteurs ont considéré deux marqueurs à la fois pour les haplotypes. Les fréquences haplotypiques ont été estimées à partir de l'algorithme EM pour les deux types d'échantillon. Cette étude est celle qui s'approche le plus de la nôtre, quoique les tests haplotypiques comparés ne soient pas les mêmes. De notre côté, nous avons choisi des tests plus courants, soit les tests du score et FBAT.

Dans les deux cas, les auteurs concluent que l'approche haplotypique n'est pas plus avantageuse que l'approche allélique. Selon eux, ceci s'expliquerait en partie par la présence d'un fort déséquilibre de liaison entre les marqueurs. Shephard et coll. (2005), qui à l'instar de Hintsanen et coll. (2006) ont aussi comparé la puissance des deux devis, concluent que les tests basés sur des cas-témoins sont plus puissants que ceux basés sur des trios, même lorsque les devis comportent un nombre de cas identique.

Chapitre 8

Description de l'étude de Monte-Carlo

Pour répondre à l'objectif principal de ce mémoire, qui est de comparer divers tests d'association génétique basés sur le déséquilibre de liaison (DL), une étude de Monte-Carlo a été réalisée afin d'évaluer la puissance statistique des devis d'échantillonnage de cas-témoins et de trios et des approches analytiques allélique et haplotypique.

Ce chapitre comporte deux parties. Les détails de la procédure de simulation sont présentés à la section 8.1 ; en plus d'y préciser les tailles d'échantillons et la façon dont on a déterminé le nombre de réplicats, on y décrit la nature de l'association que l'on a choisi de simuler entre la maladie et les marqueurs. Dans la section 8.2, on explique les démarches qui ont été employées en vue de calibrer les tests considérés. Ces derniers sont décrits aux chapitres 4 et 6, tel que précisé au tableau 8.1.

		association haplotypique	
devis	association allélique	global	spécifique
cas-témoins	χ^2 sur les allèles	TRV SCORE	TRV SCORE
trios	TDT	FBAT	FBAT

Tab. 8.1 – Les tests comparés dans le cadre de l'étude de Monte-Carlo

8.1 Détails de la procédure de simulation

Les paramètres de l'étude de Monte-Carlo ont été choisis de façon à faciliter au maximum les comparaisons de puissance.

8.1.1 Contexte de la simulation

8.1.1.1 Tailles des échantillons

Les tailles d'échantillon ont été fixées à 50 trios et à 50 cas et 50 témoins. Notons que chaque trio simulé est apparié à un couple de cas-témoin. Se rapporter à la section 8.1.3 pour de plus amples détails.

8.1.1.2 Caractéristiques du trait

Le trait choisi pour la simulation est la présence de la maladie. On a donc un trait dichotomique (0 : absence de la maladie, 1 : présence de la maladie). Le nombre d'allèles possibles au locus de la maladie a été fixé à deux. L'allèle mutant, celui qui cause la maladie, est noté D ; quant à l'allèle normal, il est noté d.

On suppose que la transmission de la maladie se fait sous un mode dominant. Autrement dit, une seule copie de l'allèle mutant suffit pour que le sujet soit affecté par la maladie. En d'autres termes, le porteur du génotype D/D ou D/d est malade et celui qui porte le génotype d/d ne l'est pas.

Dans les échantillons de trios, au moins un des parents doit être atteint de la maladie pour qu'elle ait été transmise à l'enfant ; on peut supposer sans perte de généralité qu'il s'agit de la mère. On suppose en outre que l'enfant malade est le seul qui soit atteint de la maladie dans la famille. Ainsi, le génotype de la mère doit contenir au moins une copie de l'allèle mutant D et ne peut pas en contenir deux puisqu'alors elle aurait transmis la maladie à tous ses enfants. Par conséquent, la mère doit forcément avoir le génotype D/d. Les génotypes au locus de la maladie pour les membres d'un trio sont présentés au tableau 8.2.

membre	maladie	génoytpe
père	0	d/d
mère	1	D/d
enfant	1	D/d

TAB. 8.2 – Les génotypes au locus de la maladie pour les membres d'un trio

8.1.1.3 Caractéristiques des marqueurs

Le nombre de marqueurs pour la simulation a été fixé à deux. Les marqueurs sont notés $M1$ et $M2$. Chacun possède deux allèles, numérotés 1 et 2. Quatre haplotypes sont donc possibles : 11 (lire 1, 1 et non pas onze), 12, 21 et 22.

On suppose que le locus de la maladie est situé à égale distance entre les deux marqueurs, tel qu'illustré à la figure 8.1. Cet emplacement des marqueurs par rapport au locus de la maladie entraîne l'égalité des coefficients de DL entre chacun des marqueurs et la maladie, de sorte que

$$\mathcal{D}_{M1\text{-maladie}} = \mathcal{D}_{M2\text{-maladie}}.$$

De ce fait, les marqueurs $M1$ et $M2$ ont le même niveau d'association avec la maladie.

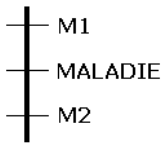


FIG. 8.1 – *Position des marqueurs $M1$ et $M2$ par rapport au locus de la maladie*

Afin de respecter l'hypothèse d'absence de recombinaison entre les marqueurs, exigée par les tests FBAT (revoir la section 6.2.2), on suppose qu'aucune recombinaison ne peut survenir entre les marqueurs. On suppose également qu'il ne peut y avoir de recombinaison entre un marqueur et la maladie. Bref, on fait l'hypothèse qu'aucune recombinaison ne peut survenir entre les trois loci.

8.1.1.4 Nature de l'association entre la maladie et les marqueurs

L'association entre la maladie et les marqueurs est simulée au niveau des haplotypes. On suppose que

- l'haplotype 11 est plus fréquent chez les cas que chez les témoins ;
- les haplotypes 12 et 21 sont aussi fréquents chez les cas que chez les témoins ;
- l'haplotype 22 est moins fréquent chez les cas que chez les témoins.

En d'autres termes,

- le rapport de cotes pour l'haplotype 11 sera supérieur à 1 ;
- le rapport de cotes pour les haplotypes 12 et 21 sera égal à 1 ;
- le rapport de cotes pour l'haplotype 22 sera inférieur à 1.

Puisque l'haplotype 11 est plus fréquent chez les sujets malades, il existe une association positive entre la maladie et les allèles 1 des deux marqueurs. La force de cette association sera par ailleurs la même pour les deux marqueurs.

8.1.2 Simuler l'association entre la maladie et les marqueurs

Le principe choisi pour simuler l'association entre la maladie et les marqueurs se résume à établir une distribution des fréquences haplotypiques spécifique à chaque allèle du locus de la maladie. Cette procédure est entre autres utilisée par le logiciel SIMLA de simulations pour les études génétiques ; pour de plus amples détails concernant ce logiciel, se référer à Bass et coll. (2004).

8.1.2.1 Distributions des fréquences haplotypiques pour les allèles du locus de la maladie

Le tableau 8.3 présente les paramètres pour les distributions des fréquences haplotypiques des allèles du locus de la maladie. Les notations p_{ij} et q_{ij} réfèrent aux probabilités conditionnelles de l'haplotype ij sachant l'allèle du locus de la maladie. Ainsi par exemple p_{11} correspond à la probabilité d'avoir l'haplotype 11 sachant que l'allèle D est présent au locus de la maladie. Notons que les fréquences haplotypiques sont toutes positives. De plus, les fréquences pour un allèle particulier somment à 1.

	allèle		
haplotype	D	d	marge
11	p_{11}	q_{11}	m_{11}
12	p_{12}	q_{12}	m_{12}
21	p_{21}	q_{21}	m_{21}
22	p_{22}	q_{22}	m_{22}
somme	1	1	1

TAB. 8.3 – Distributions des fréquences haplotypiques pour les allèles du locus de la maladie

On suppose de plus que les fréquences d'allèles du locus de la maladie sont égales, c'est-à-dire que $\mathrm{P}(D) = \mathrm{P}(d) = 0.5$. Les fréquences haplotypiques marginales m_{ij} s'obtiennent alors à partir des équations suivantes :

$$\begin{aligned} m_{11} &= 0.5 \times p_{11} + 0.5 \times q_{11}, \\ m_{12} &= 0.5 \times p_{12} + 0.5 \times q_{12}, \\ m_{21} &= 0.5 \times p_{21} + 0.5 \times q_{21}, \\ m_{22} &= 0.5 \times p_{22} + 0.5 \times q_{22}. \end{aligned}$$

8.1.2.2 Distributions des fréquences haplotypiques chez les cas et chez les témoins

On peut également s'intéresser à la distribution des fréquences haplotypiques chez les cas et chez les témoins, pour le calcul des rapports de cotes par exemple. Pour les cas, qui ont un génotype D/d au locus de la maladie, on obtient ces fréquences en combinant les fréquences haplotypiques des allèles D et d. Par exemple, la probabilité conditionnelle pour l'haplotype 11 chez les cas est la suivante :

$$\mathrm{P}(11|\mathrm{cas}) = \mathrm{P}(11|D/d) = 0.5 \times \mathrm{P}(11|D) + 0.5 \times \mathrm{P}(11|d) = 0.5 \times p_{11} + 0.5 \times q_{11} = m_{11}.$$

En fait, les fréquences haplotypiques des cas sont égales aux fréquences marginales, définies à la section 8.1.2.1. Pour les témoins, qui ont un génotype d/d au locus de la maladie, les fréquences haplotypiques sont les mêmes que celles pour l'allèle d.

Le tableau 8.4 présente les fréquences haplotypiques des deux groupes. Pour simuler l'association entre la maladie et les marqueurs décrite au paragraphe 8.1.1.4 de la section 8.1.1, on doit respecter les quatre conditions suivantes :

$$m_{11} > q_{11}, \quad m_{12} = q_{12}, \quad m_{21} = q_{21}, \quad m_{22} < q_2.$$

haplotype	cas (D/d)	témoins (d/d)
11	m_{11}	q_{11}
12	m_{12}	q_{12}
21	m_{21}	q_{21}
22	m_{22}	q_{22}
somme	1	1

TAB. 8.4 – Distributions des fréquences haplotypiques chez les cas et les témoins

On peut calculer les rapports de cotes facilement à partir du tableau 8.4. Par exemple, pour l'haplotype 11, on a

$$RC_{11} = \frac{m_{11} \times (1 - q_{11})}{q_{11} \times (1 - m_{11})} .$$

Pour s'assurer que les rapports de cotes soient évaluables, les fréquences haplotypiques des cas et des témoins doivent être non nulles.

8.1.2.3 Mesure du déséquilibre de liaison entre les marqueurs

La notation utilisée pour les fréquences alléliques des marqueurs $M1$ et $M2$ est la suivante :

$m_{1\bullet}$ = fréquence de l'allèle 1 pour le marqueur $M1$;

$m_{2\bullet}$ = fréquence de l'allèle 2 pour le marqueur $M1$;

$m_{\bullet 1}$ = fréquence de l'allèle 1 pour le marqueur $M2$;

$m_{\bullet 2}$ = fréquence de l'allèle 2 pour le marqueur $M2$.

On restreint le nombre de cas possibles pour les fréquences alléliques en choisissant des fréquences égales pour les deux marqueurs, c'est-à-dire que $m_{1\bullet} = m_{\bullet 1}$ et $m_{2\bullet} = m_{\bullet 2}$.

Le tableau 8.5 présente les fréquences haplotypiques pour les marqueurs $M1$ et $M2$. Notons que ces fréquences sont en fait les fréquences marginales du tableau 8.3. Le coefficient de DL entre les marqueurs s'obtient à partir de l'équation suivante :

$$\mathcal{D}_{M1\text{-}M2} = m_{11} - m_{1\bullet} \times m_{\bullet 1}.$$

		$M2$		
		1	2	
$M1$	1	m_{11}	m_{12}	$m_{1\bullet}$
	2	m_{21}	m_{22}	$m_{2\bullet}$
		$m_{\bullet 1}$	$m_{\bullet 2}$	1

TAB. 8.5 – Fréquences haplotypiques pour les marqueurs $M1$ et $M2$

8.1.2.4 Mesures de l'association entre la maladie et les marqueurs

L'association simulée entre la maladie et les marqueurs se mesure par les coefficients de DL entre la maladie et chacun des marqueurs. Pour les calculer, on doit avoir les fréquences des haplotypes formés par la maladie avec chaque marqueur.

Les tableaux 8.6 et 8.7 présentent ces fréquences haplotypiques, lesquelles se calculent à partir des valeurs du tableau 8.3.

		$M1$		
		1	2	
maladie	D	$(p_{11}+p_{12})/2$	$(p_{21}+p_{22})/2$	0.5
	d	$(q_{11}+q_{12})/2$	$(q_{21}+q_{22})/2$	0.5
		$m_{1\bullet}$	$m_{2\bullet}$	1

TAB. 8.6 – Fréquences haplotypiques pour la maladie et le marqueur $M1$

		$M2$		
		1	2	
maladie	D	$(p_{11}+p_{21})/2$	$(p_{12}+p_{22})/2$	0.5
	d	$(q_{11}+q_{21})/2$	$(q_{12}+q_{22})/2$	0.5
		$m_{\bullet 1}$	$m_{\bullet 2}$	1

TAB. 8.7 – Fréquences haplotypiques pour la maladie et le marqueur $M2$

Si l'on s'intéresse au DL entre l'allèle mutant D et l'allèle 1 de chaque marqueur, les coefficients de DL sont les suivants :

$$\mathcal{D}_{M1\text{-maladie}} = (p_{11}+p_{12})/2 - 0.5 \times m_{1\bullet}, \quad \mathcal{D}_{M2\text{-maladie}} = (p_{11}+p_{21})/2 - 0.5 \times m_{\bullet 1}.$$

Rappelons que les coefficients de DL entre chacun des marqueurs et la maladie doivent être égaux étant donné leur emplacement par rapport au locus de la maladie

(revoir la figure 8.1). Pour respecter cette condition, il suffit que les fréquences haplotypiques p_{12} et p_{21} soient égales, puisque $m_{1\bullet}$ et $m_{\bullet 1}$ ont été supposées les mêmes au paragraphe 8.1.2.3 de la section 8.1.2.

Lorsque le coefficient de DL a été défini à la section 3.1.2, on a vu que le calcul pouvait se faire avec l'un ou l'autre des allèles de chaque marqueur. On pourrait par exemple choisir dans ce cas-ci de considérer l'allèle normal plutôt que l'allèle mutant. Les coefficients de DL seraient alors les suivants :

$$\mathcal{D}_{M1\text{-maladie}} = (q_{11} + q_{12})/2 - 0.5 \times m_{1\bullet}, \quad \mathcal{D}_{M2\text{-maladie}} = (q_{11} + q_{21})/2 - 0.5 \times m_{\bullet 1}.$$

Pour faire en sorte que ces coefficients soient égaux, il suffit cette fois que les fréquences haplotypiques q_{12} et q_{21} soient les mêmes.

Puisque les fréquences p_{12} et p_{21} doivent être égales ainsi que les fréquences q_{12} et q_{21}, on doit également exiger que les fréquences marginales m_{12} et m_{21} soient égales (revoir les équations au paragraphe 8.1.2.1 de la section 8.1.2).

8.1.2.5 Étapes pour simuler l'association entre la maladie et les marqueurs

La procédure utilisée pour simuler l'association entre la maladie et les marqueurs se résume en quatre étapes. La figure 8.2 les illustre à l'aide d'un exemple.

Étape 1 : Fixer les fréquences alléliques des deux marqueurs en choisissant des fréquences égales ($m_{1\bullet} = m_{\bullet 1}$, $m_{2\bullet} = m_{\bullet 2}$).

Étape 2 : Fixer la valeur du coefficient $\mathcal{D}_{M1\text{-}M2}$ de DL entre les marqueurs.

Étape 3 : Choisir les fréquences haplotypiques pour $M1$ et $M2$ de telle sorte que la valeur du coefficient de DL entre les marqueurs soit la bonne.

Étape 4 : Choisir les fréquences haplotypiques pour les allèles du locus de la maladie.

Pour simuler divers niveaux d'association avec la maladie, on choisit plusieurs ensembles de valeurs. D'abord, pour simuler une absence d'association, qui correspond à l'hypothèse nulle, on doit avoir un coefficient de DL entre la maladie et les marqueurs égal à zéro. Pour ce faire, on choisit les mêmes fréquences haplotypiques pour les deux allèles. On les choisit donc égales aux fréquences marginales.

Puis, on augmente la valeur du DL par pas de 0.01 en modifiant les fréquences haplotypiques des deux allèles du locus de la maladie. Pour l'allèle D, à chaque pas, on augmente de 0.02 la fréquence de l'haplotype associé à la maladie, soit l'haplotype 11. Pour l'allèle d, on diminue de 0.02 la fréquence de cet haplotype. Notons que les fréquences des haplotypes 12 et 21 demeurent les mêmes puisqu'on suppose ces haplotypes non associés à la maladie.

La valeur maximale de DL pouvant être atteinte dépend des contraintes imposées par les paramètres fixés a priori. Le scénario simulé doit demeurer réaliste. Les fréquences haplotypiques ne doivent donc pas être négatives. De plus, tel que mentionné à la section 8.1.2, les fréquences haplotypiques des cas et des témoins doivent être non nulles pour que les rapports de cotes soient évaluables.

Étape 1 : Fixer les fréquences alléliques des marqueurs

		M2 1	M2 2	
M1	1	m_{11}	m_{12}	**0.7**
	2	m_{21}	m_{22}	**0.3**
		0.7	**0.3**	1

Étape 2 : Fixer $D_{\text{M1-M2}}$

$D_{\text{M1-M2}} = 0.12$

Étape 3 : Choisir les fréquences haplotypiques pour les marqueurs M1 et M2

$$
\begin{aligned}
m_{11} &= D_{\text{M1-M2}} + m_{1.} \times m_{.1} = 0.12 + 0.7^2 = 0.61 \\
m_{12} &= m_{1.} - m_{11} = 0.7 - 0.61 = 0.09 \\
m_{21} &= m_{12} = 0.09 \\
m_{22} &= m_{2.} - m_{21} = 0.3 - 0.09 = 0.21
\end{aligned}
\quad \Rightarrow
$$

		M2 1	M2 2	
M1	1	**0.61**	**0.09**	0.7
	2	**0.09**	**0.21**	0.3
		0.7	0.3	1

Étape 4 : Choisir les fréquences haplotypiques pour les allèles du locus de la maladie

Haplotype	Allèle D	Allèle d	
11	0.61	0.61	0.61
12	0.09	0.09	0.09
21	0.09	0.09	0.09
22	0.21	0.21	0.21
	1	1	

$\mathcal{D}_{\text{marqueurs-maladie}} = 0$

Haplotype	Allèle D	Allèle d	
11	**0.63**	**0.59**	0.61
12	0.09	0.09	0.09
21	0.09	0.09	0.09
22	**0.19**	**0.23**	0.21
	1	1	

$\mathcal{D}_{\text{marqueurs-maladie}} = 0.01$

$\ldots$

Haplotype	Allèle D	Allèle d	
11	**0.82**	**0.4**	0.61
12	0.09	0.09	0.09
21	0.09	0.09	0.09
22	**0**	**0.42**	0.21
	1	1	

$\mathcal{D}_{\text{marqueurs-maladie}} = 0.105$

FIG. 8.2 – *Illustration des étapes pour simuler l'association entre la maladie et les marqueurs à l'aide d'un exemple*

8.1.2.6 Présentation des scénarios simulés

Les paramètres fixés a priori pour les simulations sont les fréquences alléliques des marqueurs $M1$ et $M2$ et le coefficient de DL entre ces marqueurs, $\mathcal{D}_{M1\text{-}M2}$. Deux ensembles de valeurs pour les fréquences alléliques ont été considérés : un premier ensemble de valeurs avec des fréquences alléliques égales et un deuxième ensemble de valeurs avec une fréquence allélique plus grande pour l'allèle 1. Pour les coefficients de DL, trois valeurs ont été choisies. D'abord, une valeur représentant un DL faible, soit 0.05. Puis, une valeur de 0.12 qui correspond à un DL moyen. Enfin, une valeur de 0.20 qui représente un DL fort.

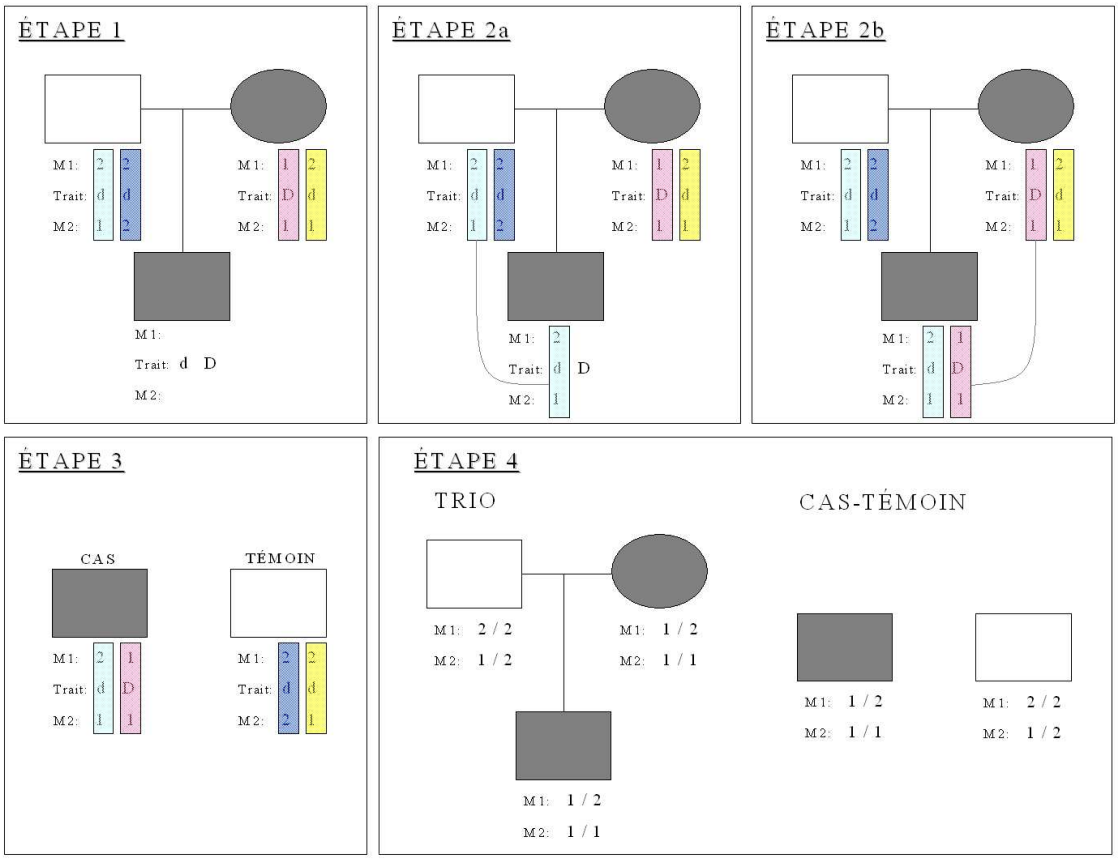
À partir de ces deux ensembles de fréquences alléliques et de ces trois valeurs de DL, six scénarios ont été simulés. Le tableau 8.8 les présente. Les différents niveaux d'association simulés pour chaque scénario apparaissent à la dernière colonne. Au total, pour les six scénarios, on compte 90 contextes différents de simulation.

		fréquences alléliques		valeurs de $\mathcal{D}_{\text{maladie-marqueurs}}$	
scénario	$\mathcal{D}_{M1-M2}$	1	2	nombre	étendue
1	0.05	0.5	0.5	15	0 – 0.14
2	0.05	0.7	0.3	8	0 – 0.07
3	0.12	0.5	0.5	16	0 – 0.15
4	0.12	0.7	0.3	12	0 – 0.105
5	0.20	0.5	0.5	23	0 – 0.22
6	0.20	0.7	0.3	15	0 – 0.145
				90	

TAB. 8.8 – Les six scénarios simulés

8.1.3 Simuler l'information génétique dans les échantillons

Pour chaque contexte de simulation, des réplicats d'échantillons de 50 trios et d'échantillons de 50 cas et 50 témoins ont été générés. La procédure de simulation a été choisie pour faire en sorte que chaque trio simulé soit apparié à un couple de cas-témoin. Par conséquent, chaque échantillon de trios est apparié à un échantillon de cas-témoins. Cette procédure de simulation se résume en quatre étapes, illustrées à la figure 8.3.



Étape 1 : simuler les haplotypes des parents selon les distributions associées à D et d.

Étape 2 :

a) simuler l'haplotype que l'enfant reçoit du père (chaque haplotype a une probabilité 1/2 d'être transmis) ;

b) attribuer à l'enfant l'haplotype transmis par la mère qui contient l'allèle de la maladie, D.

Étape 3 : les haplotypes non transmis par les parents deviennent les haplotypes du témoin dans l'échantillon de cas-témoins.

Étape 4 : ramener l'information génétique sous forme de génotype aux marqueurs.

FIG. 8.3 – *Schéma de simulation d'un trio et d'un couple de cas-témoin*

8.1.4 Détermination du nombre de réplicats

Une étude de puissance a été réalisée dans le but d'identifier le nombre de réplicats optimal pour détecter une différence notable de puissance entre deux tests. La puissance d'un test correspond à la proportion de rejet de l'hypothèse nulle parmi tous les réplicats. En pratique, une différence de puissance inférieure à 5% n'est pas suffisante pour écarter le test le moins puissant. On s'est donc intéressé à des différences de puissance variant de 5% à 10%.

Idéalement, étant donné l'appariement des échantillons de trios et de cas-témoins, on aurait utilisé la formule de calcul de taille d'échantillon pour un test de McNemar

qui permet de comparer les proportions de deux échantillons appariés. Toutefois, ce calcul nécessite la proportion de paires discordantes espérée, ce qui ne correspond pas à la valeur qui nous intéresse, c'est-à-dire à la différence entre les deux proportions. Il a donc été convenu d'utiliser la formule de calcul de taille d'échantillon pour comparer les proportions de deux échantillons indépendants de même taille.

p_1 : puissance du test n° 1	p_2 : puissance du test n° 2	tailles d'échantillon
0.30	0.35	2278
0.30	0.37	1179
0.30	0.40	589
0.40	0.45	2538
0.40	0.47	1301
0.40	0.50	641
0.50	0.55	2590
0.50	0.57	1317
0.50	0.60	641
0.60	0.65	2434
0.60	0.67	1227
0.60	0.70	589
0.70	0.75	2070
0.70	0.77	1031
0.70	0.80	485
0.80	0.85	1498
0.80	0.87	728
0.80	0.90	329
0.90	0.95	719
0.90	0.97	320
0.90	1.00	121

TAB. 8.9 – Tailles d'échantillon minimales pour détecter des différences de puissance de l'ordre de 5% à 10% ; $\pi = 95\%$; $\alpha = 5\%$

Étant donné deux proportions p_1 et p_2, il existe une formule simple permettant de déterminer la taille minimale requise pour qu'un test effectué au seuil de 5% ait 95% des chances de rejeter $H_0 : p_1 = p_2$ pour des valeurs de p_1 et p_2 données. On obtient cette taille minimale à partir de l'équation suivante :

$$n = \frac{\left\{ \mathrm{Z}_{1-\frac{\alpha}{2}} \sqrt{(p_1+p_2)\{1-(p_1+p_2)/2\}} + \mathrm{Z}_{1-\beta} \sqrt{p_1(1-p_1)+p_2(1-p_2)} \right\}^2}{(p_1-p_2)^2},$$

où en général Z_γ représente le quantile d'ordre γ d'une loi normale centrée réduite.

Ces tailles minimales d'échantillons (indépendants) sont données au tableau 8.9 pour différentes valeurs de p_1 et p_2. Ces proportions correspondent ici aux puissances des deux tests que l'on cherche à comparer.

1. Simuler** 2500 **réplicats d'échantillons appariés de 50 trios et de 50 cas et 50 témoins.

2. Conserver les** 2000 **premiers réplicats pour lesquels l'équilibre de Hardy–Weinberg est respecté pour les deux marqueurs chez les témoins des échantillons de cas-témoins.

Échantillons simulés	
TRIOS	CAS-TÉMOINS
trios_1	cas_temoins_1
trios_2	cas_temoins_2
trios_3	cas_temoins_3
⋮	⋮
trios_1001	cas_temoins_1001
trios_1002	cas_temoins_1002
trios_1003	cas_temoins_1003
⋮	⋮
trios_2498	cas_temoins_2498
trios_2499	cas_temoins_2499
trios_2500	cas_temoins_2500

Tester l'équilibre de Hardy–Weinberg chez les témoins

Échantillons simulés	
TRIOS	CAS-TÉMOINS
trios_1	cas_temoins_1
trios_2	cas_temoins_2
trios_3	cas_temoins_3
⋮	⋮
trios_1001	cas_temoins_1001
trios_1002	cas_temoins_1002
trios_1003	cas_temoins_1003
⋮	⋮
trios_2498	cas_temoins_2498
trios_2499	cas_temoins_2499
trios_2500	cas_temoins_2500

2000 réplicats

3. Pour chaque réplicat, effectuer les tests d'association dans les deux types d'échantillon :

a) dans l'échantillon de cas-témoins :

i) test allélique : test du khi-deux sur les allèles avec la procédure `casecontrol` du module `genetics` de SAS.

ii) tests haplotypiques :

- test TRV avec la procédure `haplotype` du module `genetics` de SAS ;
- test du score avec la fonction `haplo.score` du logiciel R.

b) dans l'échantillon de trios :

i) test allélique : test TDT avec le logiciel FBAT.

ii) tests haplotypiques : test FBAT avec le logiciel FBAT.

4. Calculer la puissance de chaque test :

$$\text{nombre de réplicats qui rejettent } H_0 / 2000$$

FIG. 8.4 – *Algorithme d'analyse pour chaque contexte de simulation*

En consultant le tableau 8.9, on constate que les tailles minimales d'échantillons varient de 121 à 2590 selon les puissances des deux tests. Étant donné que l'on s'intéresse plus particulièrement à des différences de puissance supérieures à 5%, le nombre de réplicats a finalement été fixé à 2000.

8.1.5 Description de l'algorithme d'analyse

Un algorithme a été implanté afin d'automatiser la simulation et l'analyse des données pour les six scénarios à l'étude. La figure 8.4 résume les étapes de l'algorithme à répéter pour chacun des 90 contextes de simulation présentés à la section 8.1.2.

8.2 Calibrage des tests

Avant de procéder à l'étude de simulation, on a d'abord évalué l'erreur de première espèce de chacun des tests afin de déterminer s'il y avait lieu de les calibrer.

On se rappelle que par définition le seuil de signification, α, représente la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie. On s'attend donc à ce que la puissance d'un test bien calibré donne une valeur qui soit très proche de α lorsque les données analysées ont été simulées sous l'hypothèse nulle. Si un test s'avère mal calibré, on peut corriger la situation en ajustant le seuil de signification ou encore la valeur critique du test. Dans le cadre de notre étude, on a opté pour un ajustement de la valeur critique lorsqu'un calibrage s'avérait nécessaire.

Comparer des tests mal calibrés peut conduire à des conclusions erronées. Supposons par exemple que deux tests soient effectués au seuil nominal de 5%. Supposons de plus que le premier test est bien calibré et que le second rejette l'hypothèse nulle dans plus de 5% des cas lorsqu'elle est vraie. Le deuxième test risque alors de paraître, à tort, plus puissant que le premier.

Pour éviter de telles situations, une évaluation de l'erreur de première espèce a été effectuée pour les tests d'association allélique ainsi que pour les tests globaux et spécifiques d'association haplotypique. Les valeurs critiques, ajustées au besoin, seront présentées pour l'ensemble des tests. Cependant, afin d'alléger la présentation de cette section, on présente les détails de l'évaluation de l'erreur de première espèce pour les tests globaux d'association haplotypique seulement.

8.2.1 Évaluation de l'erreur de première espèce pour les tests globaux d'association haplotypique

On a simulé 11 500 réplicats d'échantillons de 50 trios et d'échantillons de 50 cas et 50 témoins sous l'hypothèse nulle, c'est-à-dire sous l'hypothèse qu'aucun haplotype n'est associé à la maladie. Le contexte de simulation choisi est celui du scénario 3 avec des fréquences égales à 0.5 pour les deux allèles des marqueurs $M1$ et $M2$ et un coefficient de DL entre ces marqueurs égal à 0.12. L'équilibre de Hardy–Weinberg a été testé chez les témoins des échantillons de cas-témoins.

Pour s'assurer que l'interprétation des tests soit valide, on a considéré les 10 183 réplicats pour lesquels l'équilibre de Hardy–Weinberg était respecté pour les deux marqueurs. Le seuil de signification des tests a été fixé à 5%. Par conséquent, la puissance attendue était de 5% sous H_0. La puisssance observée pour chacun des tests, soit l'erreur de première espèce, est présentée au tableau 8.10.

test	nombre de réplicats	erreur de première espèce
SCORE	10 183	0.0506
TRV	10 183	0.0557
FBAT	10 183	0.0461

Tab. 8.10 – Erreur de première espèce pour les trois tests globaux d'association haplotypique

On remarque au départ que l'erreur de première espèce des trois tests est relativement proche de 5%, ce qui laisse à penser qu'ils sont bien calibrés. Avant de conclure que tel est le cas, on a cependant cherché à déterminer si la distribution des statistiques de chacun des tests concordait bien avec leur distribution asymptotique sous l'hypothèse nulle. Pour ce faire, on a effectué des tests d'ajustement de Kolmogorov–Smirnov. Par la suite, on a évalué la précision de la queue de la distribution de chacun des tests. Enfin, on a effectué des tests de McNemar pour évaluer le niveau de discordance entre les tests.

8.2.1.1 Tests d'ajustement de Kolmogorov–Smirnov

Dans le cadre de cette étude, le nombre d'haplotypes possible est quatre. Sous l'hypothèse nulle, la distribution asymptotique des trois tests est une khi-deux à trois

degrés de liberté (revoir les sections 4.2.2, 4.2.3 et 6.2.1 au besoin).

Des tests de Kolmogorov–Smirnov ont été effectués dans le but de déterminer si la distribution empirique des statistiques de chaque test s'ajuste bien à la distribution théorique d'une khi-deux à trois degrés de liberté. En plus du test de Kolmogorov–Smirnov, on peut également supposer que la loi des observations est une $\Gamma(\gamma, 2)$ et estimer le paramètre γ afin de voir s'il se rapproche de 1.5, dans lequel cas la loi serait bien une khi-deux à trois degrés de liberté.

test	statistique de K-S	seuil observé	$\hat{\gamma}$
SCORE	0.0064	> 0.25	1.5076
TRV	0.0086	> 0.25	1.5085
FBAT	0.0082	> 0.25	1.4972

TAB. 8.11 – Tests d'ajustement de Kolmogorov–Smirnov (K-S) pour les trois tests globaux d'association haplotypique

Le tableau 8.11 présente les résultats des tests de Kolmogorov–Smirnov. La dernière colonne donne une estimation pour le paramètre γ. La figure 8.5 illustre graphiquement l'ajustement des distributions empiriques pour les trois tests par rapport à la distribution théorique d'une khi-deux à trois degrés de liberté.

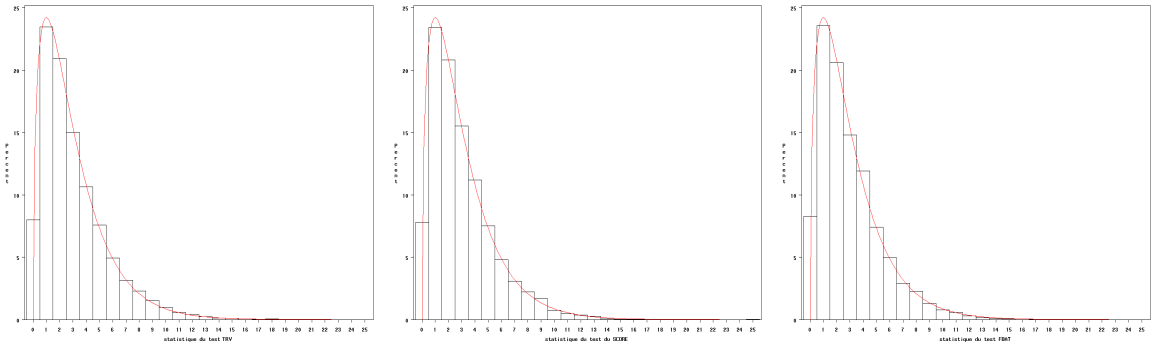


FIG. 8.5 – *Distribution des statistiques pour les trois tests globaux d'association haplotypique ; en rouge, courbe de distribution d'une khi-deux à trois degrés de liberté*

Les tests de Kolmogorov–Smirnov indiquent que les distributions empiriques des trois tests ne sont pas significativement différentes de celle d'une khi-deux à trois degrés de liberté. De plus, les estimations pour le paramètre γ sont très proches de 1.5. La

figure 8.5 permet également de conclure que les distributions empiriques des trois tests s'ajustent bien à la distribution d'une khi-deux à trois degrés de liberté.

8.2.1.2 Évaluation de la précision de la queue de la distribution

Vu l'importance de la queue de la distribution dans le cadre d'études de puissance, des tests d'égalité de proportions ont été effectués aux seuils de 1%, 5% et 10%.

Pour les données simulées sous l'hypothèse nulle, on calcule l'erreur de première espèce de chacun des tests aux trois seuils choisis. Notons que les valeurs obtenues pour un seuil de 5% ont déjà été présentées au tableau 8.10. Les erreurs de première espèce espérées sont respectivement 1%, 5% et 10%. Pour savoir si les valeurs observées sont égales à celles espérées, on effectue des tests d'égalité de proportions.

Soit X, le nombre de fois que l'hypothèse nulle est rejetée. Sa distribution est alors

$$X \sim \mathcal{BIN}(N, p),$$

où N correspond au nombre de réplicats, 10 183, et p représente la proportion théorique. On veut tester si la proportion de rejets de l'hypothèse nulle, $\hat{p}$, est significativement différente de la valeur $p = 5\%$ correspondant au seuil nominal du test. Puisque N est très grand, on peut utiliser l'approximation normale. La statistique du test est donc la suivante :

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/N}}.$$

On rejette H_0 si $|z| > Z_{1-\alpha/2} = 1.96$.

	$p = 0.01$		$p = 0.05$		$p = 0.10$	
test	$\hat{p}$	z	$\hat{p}$	z	$\hat{p}$	z
SCORE	0.0082	−1.83	0.0506	0.28	0.0994	−0.20
TRV	0.0109	0.91	0.0557	2.64	0.1064	2.15
FBAT	0.0069	−3.14	0.0461	−1.81	0.0947	−1.78

Tab. 8.12 – Tests d'égalité des proportions

Le tableau 8.12 présente les résultats pour les tests d'égalité des proportions. Pour le test du SCORE, les erreurs de première espèce ne sont pas significativement différentes des valeurs théoriques. Pour ce qui est du test TRV, les erreurs de première espèce sont

significativement différentes des valeurs théoriques lorsque le seuil de signification vaut 5% ou 10%. Enfin, on remarque que l'erreur de première espèce pour le test FBAT est différente de la valeur théorique lorsque le seuil de signification est de 1%. Ces observations laissent entendre qu'après tout, un calibrage des tests pourrait s'avérer nécessaire.

8.2.1.3 Évaluation du niveau de discordance entre les tests

Pour comparer les erreurs de première espèce des trois procédures à l'étude, des tests de McNemar ont été effectués. On aurait également pu utiliser des tests de comparaison de proportions d'échantillons indépendants. Toutefois, étant donné la nature appariée des tests, le test de McNemar s'avère un outil plus puissant. Celui-ci permet d'évaluer le niveau de discordance entre les tests. Ainsi, par exemple, on pourrait détecter deux tests qui ont des erreurs de première espèce très comparables, mais qui n'ont pas rejeté l'hypothèse nulle dans les mêmes échantillons.

Le tableau 8.13 présente les fréquences des différentes situations de concordance et de discordance entre les trois tests pour le rejet de l'hypothèse nulle. À partir de ce tableau, on peut construire trois tableaux de dimension 2×2 qui permettent d'effectuer des tests de McNemar pour comparer deux tests à la fois.

Hypothèse retenue			
test SCORE	test TRV	test FBAT	fréquence
H_0	H_0	H_0	9463
H_0	H_0	H_1	93
H_0	H_1	H_0	68
H_0	H_1	H_1	44
H_1	H_0	H_0	40
H_1	H_0	H_1	20
H_1	H_1	H_0	143
H_1	H_1	H_1	312
			10 183

TAB. 8.13 – Concordance entre les trois tests sous l'hypothèse nulle

		test du SCORE		
		H_0	H_1	
test FBAT	H_0	9531	183	9714
	H_1	137	332	469
		9668	515	10 183

TAB. 8.14 – Concordance entre les tests du SCORE et FBAT sous l'hypothèse nulle

Par exemple, le tableau 8.14 permet de comparer les tests du SCORE et FBAT. On compte au total 320 $(183+137)$ discordances entre les deux tests. La statistique du test de McNemar est donc la suivante :

$$\chi^2_{\text{McNemar}} = \frac{(183-137)^2}{183+137} = 6.61.$$

Le tableau 8.15 présente les résultats aux tests de McNemar pour la comparaison des trois tests. Au seuil de 5%, on conclut que les tests sont tous discordants entre eux, surtout les tests TRV et FBAT, ce qui confirme le besoin de calibrage.

tests comparés	χ^2_{McNemar}	seuil observé
SCORE, TRV	15.72	7×10^{-5}
SCORE, FBAT	6.61	0.0101
TRV, FBAT	29.64	5×10^{-8}

TAB. 8.15 – Résultats aux tests de McNemar pour la comparaison des trois tests

8.2.1.4 Valeurs critiques des tests globaux d'association haplotypique

L'évaluation de l'erreur de première espèce qui s'est faite en trois étapes suggère un besoin de calibrage des trois tests globaux d'association haplotypique. On a vu à la section 8.2.1.1 que les distributions empiriques ne sont pas statistiquement différentes des distributions théoriques. Cependant, la queue des distributions des tests TRV et FBAT manque de précision (revoir la section 8.2.1.2) et il existe un niveau de discordance significatif entre les tests (revoir la section 8.2.1.3). Étant donné l'objectif de cette étude qui est de comparer la puissance entre les tests, il s'avère important de s'assurer que le seuil de signification représente exactement le même pourcentage d'erreur de première espèce pour les trois tests. On évite ainsi qu'une légère supériorité de puissance soit due à une légère supériorité de l'erreur de première espèce.

Pour calibrer les tests, on a opté pour une modification de la valeur critique. Le 95^{e} percentile d'une khi-deux à trois de degrés de liberté vaut 7.8147. Cette valeur a été remplacée pour chaque test par le 95^{e} percentile de la distribution empirique de la statistique correspondante. Pour les trois tests globaux d'association haplotypique, on a obtenu les nouvelles valeurs critiques suivantes :

- Test du SCORE : 7.8564 ;
- Test TRV : 8.0544 ;
- Test FBAT : 7.6875.

Comme on l'avait remarqué à la section 8.2.1, le test du SCORE est un test bien calibré. En effet, sa nouvelle valeur critique est très proche de la valeur critique théorique. Pour le test TRV, on a maintenant une valeur critique supérieure à la valeur théorique, ce qui corrigera le fait que ce test avait tendance à rejeter l'hypothèse nulle plus souvent qu'il ne l'aurait dû. Le test FBAT, quant à lui, ne rejettait pas assez souvent l'hypothèse nulle puisque son erreur de première espèce était inférieure à 5%. La nouvelle valeur critique permettra de corriger cette situation puisqu'elle est inférieure à la valeur théorique.

Hypothèse retenue			
test SCORE	test TRV	test FBAT	nombre
H_0	H_0	H_0	9472
H_0	H_0	H_1	122
H_0	H_1	H_0	40
H_0	H_1	H_1	40
H_1	H_0	H_0	47
H_1	H_0	H_1	32
H_1	H_1	H_0	115
H_1	H_1	H_1	315
			10 183

TAB. 8.16 – Concordance entre les trois tests sous l'hypothèse nulle, à la suite du calibrage

Les tests ont par la suite été repris en considérant ces nouvelles valeurs critiques. Les erreurs de première espèce après le calibrage des tests du SCORE, TRV et FBAT sont toutes de 5%, à une erreur d'arrondi près. Le tableau 8.16 présente les fréquences

des différentes situations de concordance et de discordance entre les trois tests pour le rejet de l'hypothèse nulle à la suite du calibrage.

En comparant les tableaux 8.13 et 8.16, on constate que le calibrage a permis de diminuer le niveau de discordance entre les tests. Pour s'en assurer, il convient de reprendre les tests de McNemar pour comparer les procédures deux à la fois. Le tableau 8.17 présente ces résultats. La discordance entre les tests n'est maintenant plus significative. Le calibrage permettra une meilleure comparaison des tests.

tests comparés	χ^2_{McNemar}	seuil observé
SCORE, TRV	0.0063	0.9367
SCORE, FBAT	0	1
TRV, FBAT	0.0032	0.9549

TAB. 8.17 – Résultats aux tests de McNemar pour la comparaison des trois tests, à la suite du calibrage

8.2.2 Nouvelles valeurs critiques pour tous les tests

La procédure utilisée pour déterminer si le calibrage des tests globaux d'association hapotypique était nécessaire a aussi été appliquée aux autres procédures à partir du même échantillon de 10 183 réplicats. Tous ont finalement été calibrés. Le tableau 8.18 résume les nouvelles valeurs critiques de l'ensemble des tests d'association qui font l'objet de l'étude de simulation.

	association allélique		association haplotypique		
				valeur critique	
devis	test	valeur critique	test	global	spécifique
cas-témoins	χ^2	3.9276	SCORE	7.8564	3.7958
			TRV	8.0544	4.2179
trios	TDT	3.919996	FBAT	7.6875	3.7779

TAB. 8.18 – Récapitulation des valeurs critiques après le calibrage des tests

Chapitre 9

Résultats et discussion

Ce chapitre est consacré à la présentation et à l'interprétation des résultats de l'étude de Monte-Carlo. La puissance des devis d'échantillonnage de cas-témoins et de trios est comparée dans la section 9.1 ; la différence de puissance entre les tests alléliques et haplotypiques est quantifiée à la section 9.2. Un résumé des conclusions et une discussion générale se trouvent dans la section 9.3.

9.1 Comparaison de la puissance des deux devis

La comparaison de la puissance des devis d'échantillonnage se fait en deux temps. D'abord, les courbes de puissance de chacun des tests sont présentées de telle façon que les devis d'échantillonnage puissent être comparés. Puis, on présente une évaluation de l'importance des différences de puissance, à l'aide de statistiques descriptives et de tests permettant de déterminer si les différences maximales sont significatives.

9.1.1 Courbes de puissance

La figure 9.1 illustre la distribution empirique des statistiques des trois tests globaux d'association haplotypique obtenues pour le scénario 3 décrit au tableau 8.8. L'idée est de permettre de visualiser la progression de la proportion de rejet de l'hypothèse nulle en fonction du coefficient de déséquilibre de liaison entre la maladie et les marqueurs.

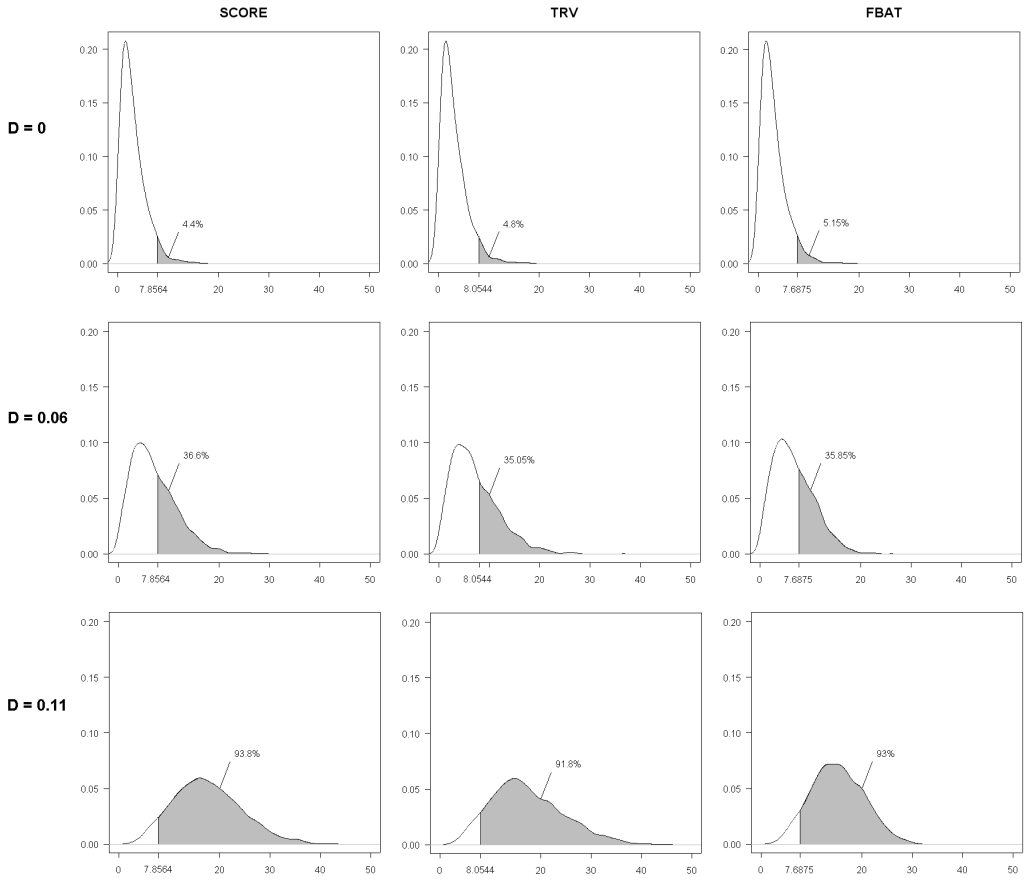


FIG. 9.1 – *Proportion de rejet de l'hypothèse nulle en fonction du coefficient de déséquilibre de liaison entre la maladie et les marqueurs ($\mathcal{D}$ = 0, 0.06, 0.11) pour les trois tests globaux d'association haplotypique ; scénario 3*

Les trois figures suivantes présentent les courbes de puissance des tests d'association allélique (9.2), des tests globaux d'association haplotypique (9.3) et des tests spécifiques d'association haplotypique (9.4) en fonction du déséquilibre de liaison entre la maladie et les marqueurs pour les six scénarios étudiés.

Il semble à première vue que les puissances des tests soient équivalentes pour les deux types de devis, tant pour les tests d'association allélique que haplotypique. Il appert toutefois que le test TRV est moins performant que les deux autres, SCORE et FBAT, dans le cadre de tests visant à identifier spécifiquement l'haplotype associé à la maladie.

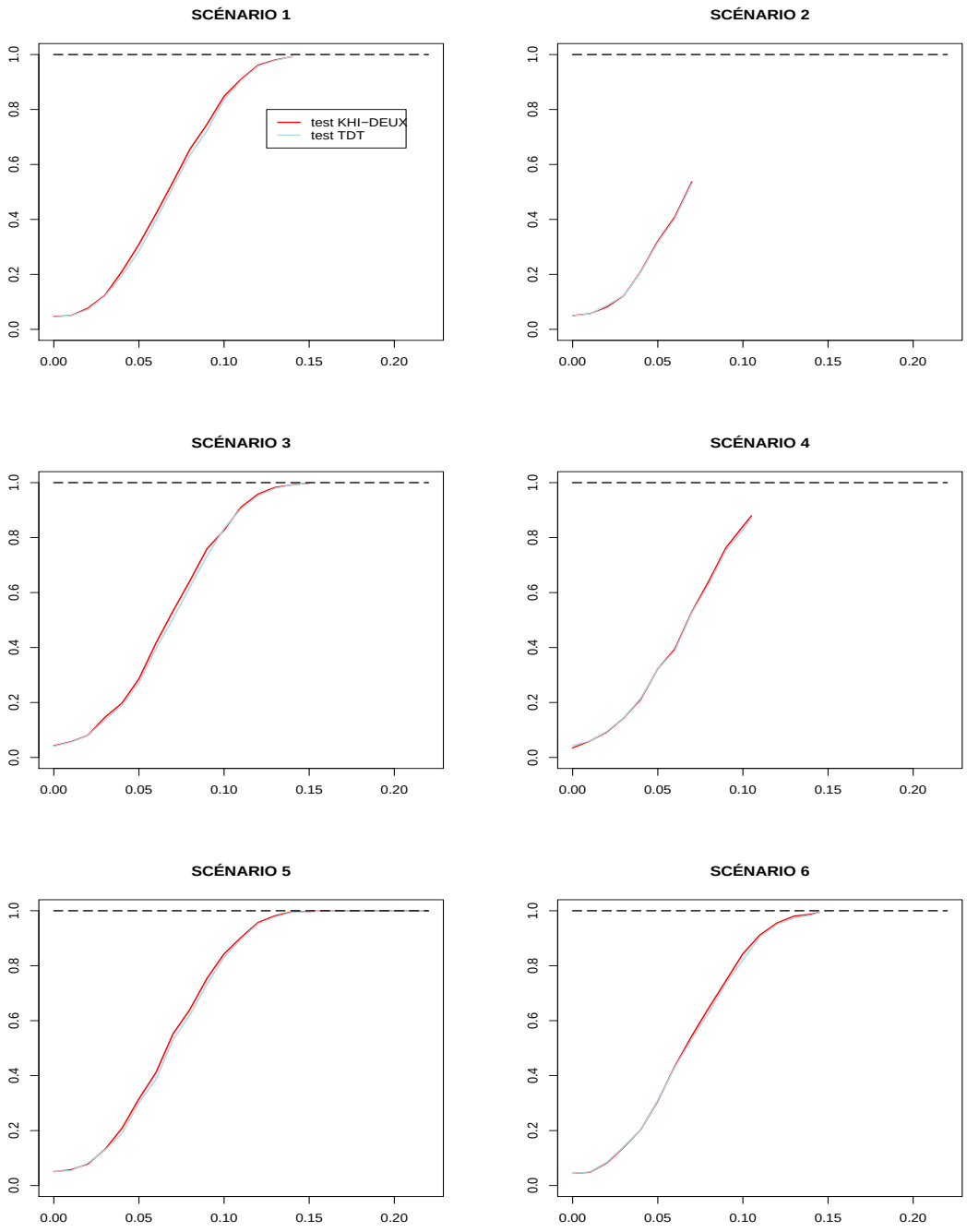


FIG. 9.2 – *Puissance des tests d'association allélique en fonction du DL entre la maladie et les marqueurs pour les six scénarios simulés ; devis cas-témoins : KHI-DEUX ; devis trios : TDT*

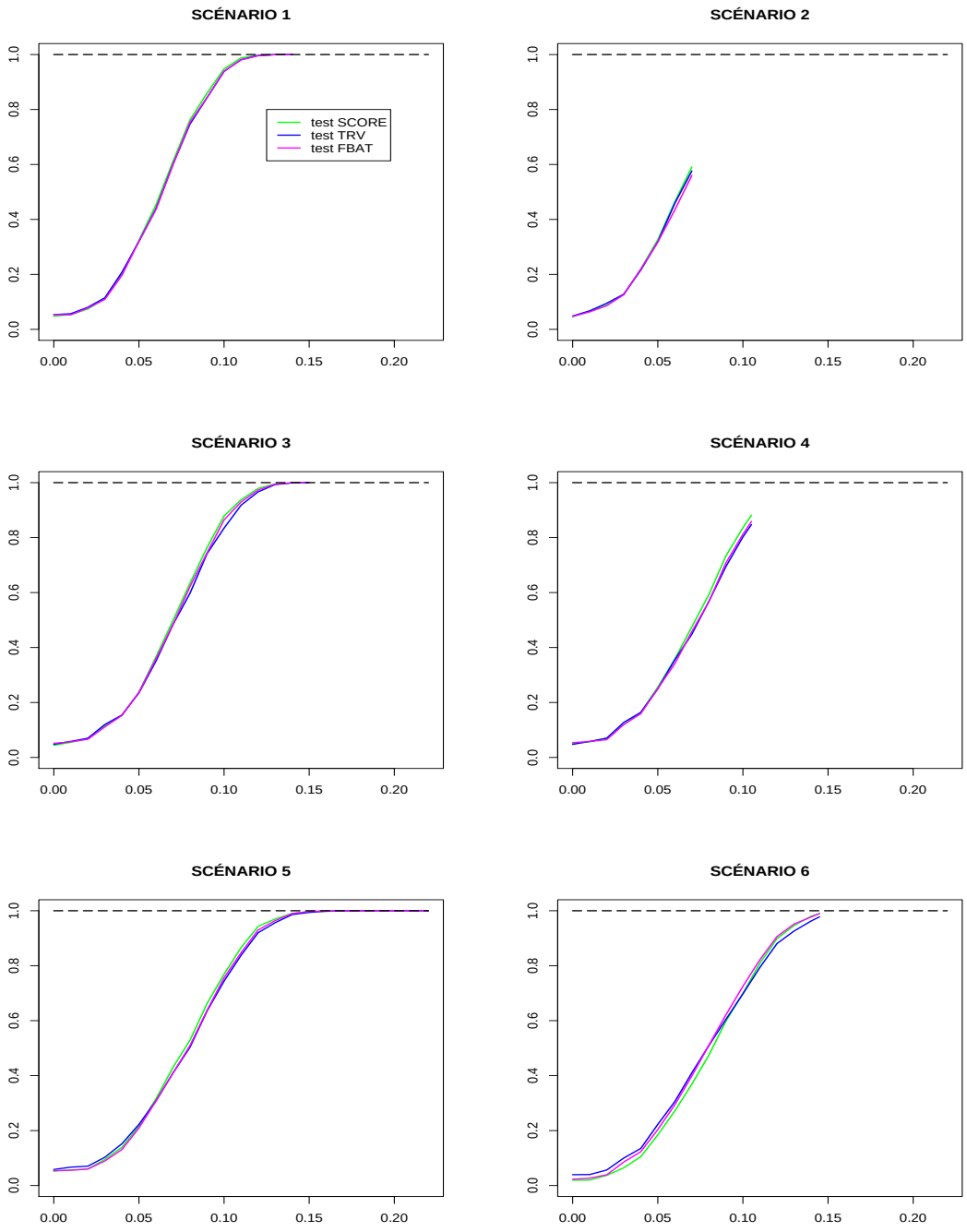


FIG. 9.3 – *Puissance des tests globaux d'association haplotypique en fonction du DL entre la maladie et les marqueurs pour les six scénarios simulés ; devis cas-témoins : SCORE et TRV ; devis trios : FBAT*

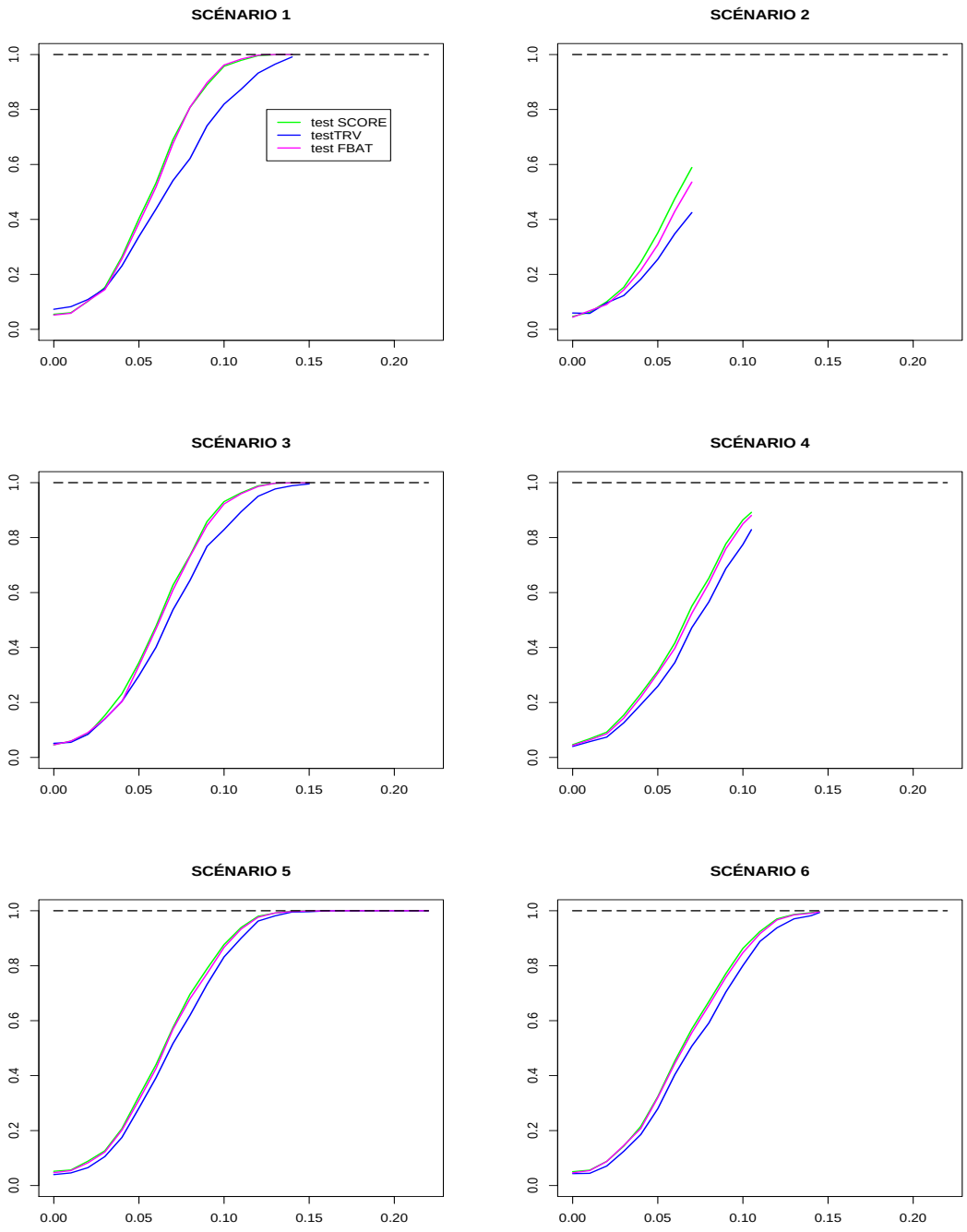


FIG. 9.4 – *Puissance des tests spécifiques d'association haplotypique en fonction du DL entre la maladie et les marqueurs pour les six scénarios simulés ; devis cas-témoins : SCORE et TRV ; devis trios : FBAT*

9.1.2 Évaluation de l'importance des différences de puissance entre les deux types de devis d'échantillonnage

Le tableau 9.1 présente les moyennes des différences de puissance entre les tests effectués dans les échantillons de cas-témoins et de trios pour les tests d'association allélique, ainsi que pour les tests globaux et spécifiques d'association haplotypique.

	moyenne de différence de puissance[a]				
	tests alléliques	tests haplotypiques globaux		tests haplotypiques spécifiques	
scénario	KHI-DEUX vs TDT	SCORE vs FBAT	TRV vs FBAT	SCORE vs FBAT	TRV vs FBAT
1	0.0105	0.0060	0.0027	0.0058	0.0699
2	0.0032	0.0098	0.0076	0.0248	0.0413
3	0.0096	0.0063	0.0067	0.0073	0.0367
4	0.0056	0.0128	0.0072	0.0127	0.0403
5	0.0070	0.0065	0.0057	0.0049	0.0169
6	0.0061	0.0148	0.0166	0.0063	0.0282

[a] En valeur absolue.

TAB. 9.1 – Moyennes de différences de puissance entre les tests effectués sur des échantillons de cas-témoins et de trios pour les approches analytiques allélique et haplotypique

On constate que les moyennes des différences de puissance ne sont pas très importantes. De fait, elles sont généralement inférieures à 2%. Les plus grandes différences sont observées au niveau des tests spécifiques d'association haplotypique, plus particulièrement entre le test TRV et FBAT, ce qui renforce l'hypothèse que le test TRV est moins puissant.

Afin de déterminer si les différences de puissance sont statistiquement significatives, des tests de McNemar ont été effectués. On s'est intéressé aux différences de puissance maximales. Les tableaux 9.2, 9.3 et 9.4 présentent respectivement les différences de puissance maximales des six scénarios ainsi que les résultats aux tests de McNemar pour les tests d'association allélique, les tests globaux d'association haplotypique et les tests spécifiques d'association haplotypique.

On conclut que les différences de puissance maximales sont généralement significatives, bien qu'elles ne soient pas très importantes. Elles sont pour la plupart inférieures à 5%, sauf entre les tests TRV et FBAT.

scénario	$\mathcal{D}_{\text{maladie-marqueurs}}$	puissance des tests		différence de puissance maximale[a]	seuil observé du test de McNemar
		KHI-DEUX	TDT		
1	0.05	0.3095	0.2845	0.0250	1.5×10^{-8}
2	0.06	0.5370	0.5315	0.0060	0.1590
3	0.07	0.5325	0.5050	0.0275	1.6×10^{-9}
4	0.10	0.8410	0.8275	0.0135	5.7×10^{-5}
5	0.06	0.4110	0.3865	0.0245	7.5×10^{-8}
6	0.10	0.8435	0.8205	0.0230	3.8×10^{-8}

[a] En valeur absolue.

TAB. 9.2 – Différences de puissance maximales entre les tests d'association allélique, KHI-DEUX vs TDT, et résultats aux tests de McNemar

scénario	$\mathcal{D}_{\text{maladie-marqueurs}}$	puissance des tests		différence de puissance maximale[a]	seuil observé du test de McNemar
		SCORE	FBAT		
1	0.06	0.4530	0.4360	0.0170	0.0122
2	0.07	0.5910	0.5600	0.0310	1.3×10^{-5}
3	0.09	0.7640	0.7445	0.0195	0.0009
4	0.08	0.5925	0.5665	0.0260	0.0003
5	0.09	0.6620	0.6375	0.0245	3.2×10^{-5}
6	0.08	0.4745	0.5105	0.0360	9.4×10^{-8}
		TRV	FBAT		
1	0.04	0.2075	0.1965	0.011	0.0858
2	0.06	0.4585	0.4345	0.024	0.0014
3	0.10	0.8345	0.8645	0.030	8.8×10^{-5}
4	0.06	0.3555	0.3415	0.014	0.0328
5	0.04	0.1520	0.1310	0.021	0.0002
6	0.11	0.7940	0.8220	0.028	6.4×10^{-6}

[a] En valeur absolue.

TAB. 9.3 – Différences de puissance maximales entre les tests globaux d'association haplotypique, SCORE vs FBAT et TRV vs FBAT, et résultats aux tests de McNemar

En ce qui concerne les tests d'association allélique, celui qui est effectué dans les échantillons de cas-témoins (le test du KHI-DEUX) est un peu plus puissant que celui effectué dans les échantillons de trios (le test TDT). De même, les tests d'association haplotypique effectués dans les échantillons de cas-témoins (les tests du SCORE et TRV) sont un peu plus puissants que celui effectué dans les échantillons de trios (le

scénario	$\mathcal{D}_{\text{maladie-marqueurs}}$	puissance des tests		différence de puissance maximale[a]	seuil observé du test de McNemar
		SCORE	FBAT		
1	0.05	0.4030	0.3870	0.0160	0.0280
2	0.07	0.5890	0.5360	0.0530	2.8×10^{-15}
3	0.04	0.2315	0.2030	0.0285	1.4×10^{-7}
4	0.07	0.5505	0.5245	0.0260	3×10^{-6}
5	0.09	0.7890	0.7715	0.0175	1.9×10^{-5}
6	0.10	0.8630	0.8475	0.0155	0.0002
		TRV	FBAT		
1	0.08	0.6215	0.8085	0.1870	8.5×10^{-71}
2	0.07	0.4250	0.5360	0.1110	2.3×10^{-37}
3	0.10	0.8290	0.9225	0.0935	2.5×10^{-30}
4	0.10	0.7750	0.8500	0.0750	5.7×10^{-31}
5	0.08	0.6195	0.6800	0.0605	2.2×10^{-20}
6	0.08	0.5915	0.6560	0.0645	7.4×10^{-25}

[a] En valeur absolue.

Tab. 9.4 – Différences de puissance maximales entre les tests spécifiques d'association haplotypique, SCORE vs FBAT et TRV vs FBAT, et résultats aux tests de McNemar

test FBAT). Toutefois, on remarque qu'au niveau des tests spécifiques d'association haplotypique, le test FBAT est plus puissant que le test TRV. Ce résultat est le plus important, puisque c'est dans ce cas que les différences de puissance sont les plus grandes et les seuils observés des tests de McNemar s'avèrent les plus petits.

9.2 Comparaison de la puissance des deux approches analytiques

La comparaison de la puissance des approches analytiques se fait elle aussi en deux temps. D'abord, les courbes de puissance de chacun des tests sont présentées de telle façon que les approches analytiques puissent être comparées. Puis, on présente une évaluation de l'importance des différences de puissance, à l'aide de statistiques descriptives et de tests permettant de déterminer si les différences maximales sont significatives.

Précisons que les résultats des tests d'association allélique ne sont présentés que pour le marqueur $M1$, ceci afin d'éviter toute redondance. En effet, la force d'association simulée est la même pour les deux marqueurs (revoir la section 8.1.1).

9.2.1 Courbes de puissance

Les figures 9.5, 9.6 et 9.7 présentent respectivement la puissance des tests d'association allélique et haplotypique des deux devis d'échantillonnage pour les scénarios impliquant un DL entre les marqueurs faible (1 et 2), moyen (3 et 4) et fort (5 et 6).

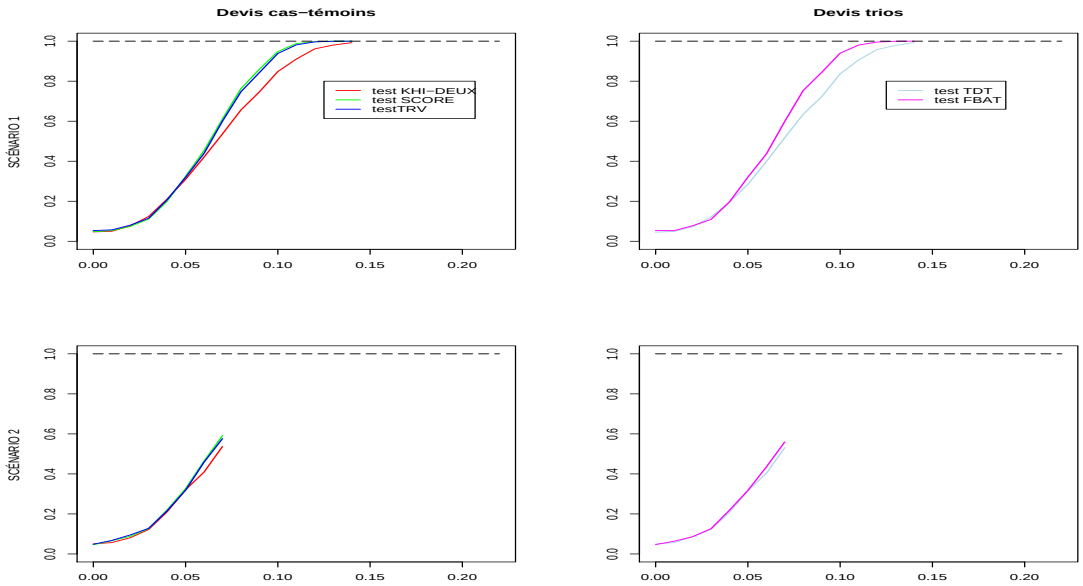


FIG. 9.5 – *Puissance des tests d'association allélique et haplotypique pour les deux devis d'échantillonnage ; tests alléliques : KHI-DEUX et TDT ; tests haplotypiques globaux : TRV, SCORE et FBAT ; scénarios 1 et 2*

On remarque que les tests alléliques sont parfois plus puissants que les tests haplotypiques, en présence d'un fort DL entre les marqueurs (scénarios 5 et 6) ou d'une plus faible association avec la maladie (scénarios 3 et 4). Les tests d'association haplotypique sont généralement avantageux si le DL entre les marqueurs est faible (scénarios 1 et 2).

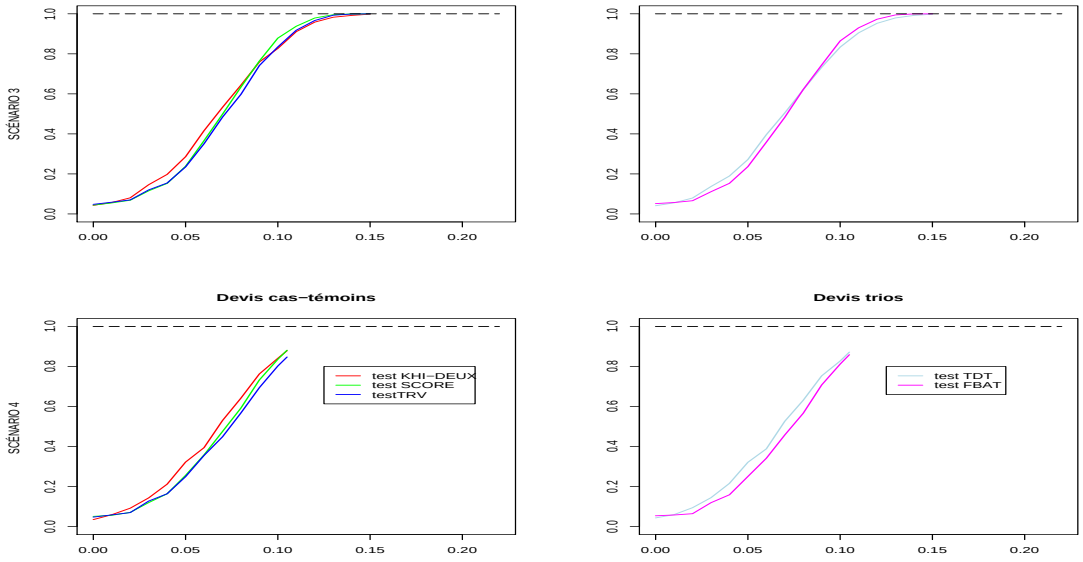


FIG. 9.6 – *Puissance des tests d'association allélique et haplotypique pour les deux devis d'échantillonnage ; tests alléliques : KHI-DEUX et TDT ; tests haplotypiques globaux : TRV, SCORE et FBAT ; scénarios 3 et 4*

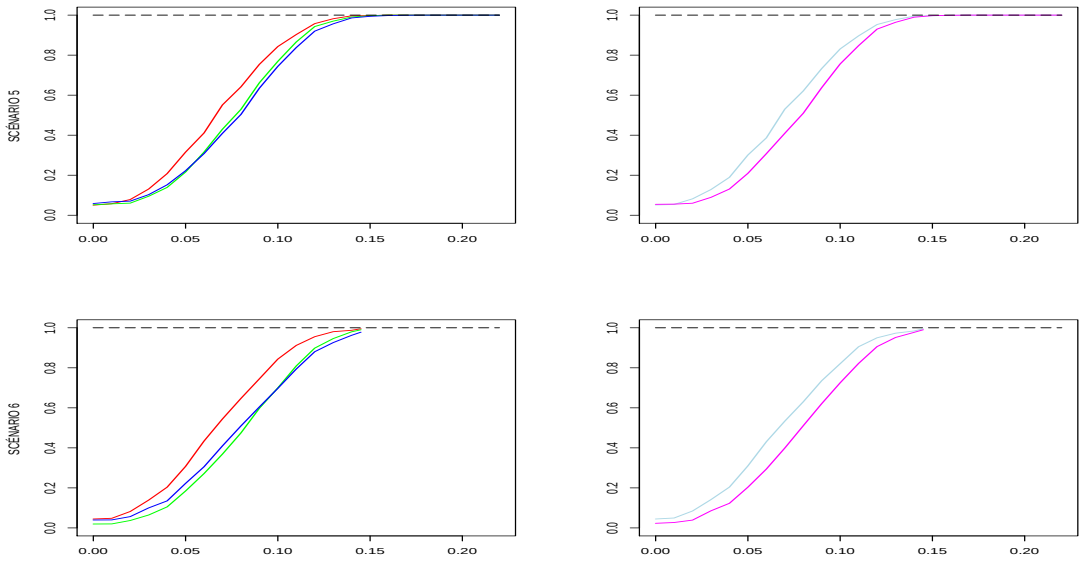


FIG. 9.7 – *Puissance des tests d'association allélique et haplotypique pour les deux devis d'échantillonnage ; tests alléliques : KHI-DEUX et TDT ; tests haplotypiques globaux : TRV, SCORE et FBAT ; scénarios 5 et 6*

9.2.2 Évaluation de l'importance des différences de puissance entre les deux approches analytiques

Le tableau 9.5 présente les moyennes de différences de puissance entre les tests d'association allélique et haplotypique pour les devis d'échantillonnage de cas-témoins et de trios.

	moyenne de différence de puissance[a]		
	devis de trios	devis de cas-témoins	
scénario	TDT vs FBAT	KHI-DEUX vs SCORE	KHI-DEUX vs TRV
1	0.0445	0.0408	0.0355
2	0.0106	0.0186	0.0160
3	0.0185	0.0223	0.0223
4	0.0375	0.0297	0.0421
5	0.0343	0.0343	0.0412
6	0.0684	0.0876	0.0755

[a] En valeur absolue.

TAB. 9.5 – Moyennes de différences de puissance entre les tests d'association allélique et haplotypique pour les devis de cas-témoins et de trios

On constate que les moyennes de différences de puissance sont plus importantes que celles observées lors de la comparaison des devis d'échantillonnage. Toutefois, elles ne dépassent pas 10%. On remarque que pour chacun des scénarios, les moyennes de différences des devis d'échantillonnage sont très comparables. On pouvait s'y attendre puisqu'à la section 9.1.1, on a noté que les puissances des deux devis étaient similaires.

Afin de déterminer si les moyennes de différences de puissance sont statistiquement significatives, des tests du khi-deux ont été effectués. Idéalement, des tests de McNemar auraient été faits, mais les fréquences des différentes situations de concordance et de discordance entre les tests n'ont malheureusement pas été compilées.

On s'est donc plutôt intéressé aux différences de puissance maximales. Les tableaux 9.6 et 9.7 présentent respectivement les différences de puissance maximales des six scénarios ainsi que les résultats aux tests du khi-deux pour les devis de trios et de cas-témoins.

scénario	$\mathcal{D}_{\text{maladie-marqueurs}}$	puissance des tests		différence de puissance maximale[a]	seuil observé du test du khi-deux
		TDT	FBAT		
1	0.09	0.7230	0.8440	0.1210	1.5×10^{-20}
2	0.06	0.4030	0.4345	0.0315	0.0435
3	0.06	0.3960	0.3585	0.0375	0.0144
4	0.05	0.3215	0.2505	0.0710	6.7×10^{-7}
5	0.07	0.5300	0.4095	0.1205	2.3×10^{-14}
6	0.06	0.4305	0.2940	0.1365	2.7×10^{-19}

[a] En valeur absolue.

Tab. 9.6 – Différences de puissance maximales entre les tests d'association allélique et haplotypique pour le devis de trios, TDT vs FBAT, et résultats aux tests du khi-deux

scénario	$\mathcal{D}_{\text{maladie-marqueurs}}$	puissance des tests		différence de puissance maximale[a]	seuil observé du test du khi-deux
		KHI-DEUX	SCORE		
1	0.09	0.7465	0.8595	0.1130	2.6×10^{-19}
2	0.06	0.4090	0.4640	0.0550	0.0005
3	0.10	0.8270	0.8785	0.0515	4.3×10^{-6}
4	0.05	0.3225	0.2560	0.0665	3.5×10^{-6}
5	0.07	0.5515	0.4300	0.1215	1.5×10^{-14}
6	0.07	0.5445	0.3690	0.1755	7.9×10^{-29}
		KHI-DEUX	TRV		
1	0.09	0.7465	0.8430	0.0965	4.2×10^{-14}
2	0.06	0.4090	0.4585	0.0495	0.0016
3	0.06	0.4160	0.3505	0.0655	2×10^{-5}
4	0.07	0.5300	0.4485	0.0815	2.5×10^{-7}
5	0.07	0.5515	0.4095	0.1420	2.5×10^{-19}
6	0.10	0.8435	0.6975	0.1460	4.8×10^{-28}

[a] En valeur absolue.

Tab. 9.7 – Différences de puissance maximales entre les tests d'association allélique et haplotypique pour le devis de cas-témoins, KHI-DEUX vs SCORE et KHI-DEUX vs TRV, et résultats aux tests du khi-deux

On conclut que les différences de puissance maximales sont généralement significatives. Elles sont pour la plupart entre 5% et 15%. Les différences les plus importantes sont observées pour les scénarios 1, 5 et 6. On constate que pour les scénarios 1 et 2, les tests d'association haplotypique sont en moyenne plus puissants que les tests d'association allélique. Pour les scénarios 3 et 4, les tests d'association allélique sont en moyenne plus puissants que les tests d'association haplotypique, avec des différences de puissance moins importantes que pour les quatre autres scénarios. Enfin, on remarque que les tests d'association allélique sont en moyenne plus puissants que les tests d'association haplotypique pour les scénarios 5 et 6.

9.3 Discussion des résultats

9.3.1 Résumé des observations

L'examen des courbes de puissance conduit aux observations suivantes. En ce qui a trait au type de devis d'échantillonnage, on remarque que :

a) même si parfois les différences sont statistiquement significatives, les puissances des tests d'association sont comparables dans les deux types de devis étudiés, et ce pour les deux approches analytiques ;
b) dans le cadre de tests spécifiques d'association haplotypique, le TRV est moins puissant que les tests du SCORE et FBAT.

En ce qui concerne l'approche analytique, on constate que :

c) la puissance des tests d'association dépend du niveau de DL entre les marqueurs ainsi que de la force de l'association avec la maladie, et ce peu importe le type de devis échantillonné ;
d) en présence d'un DL faible entre les marqueurs, l'approche analytique basée sur les haplotypes est plus puissante que celle basée sur les allèles ;
e) un fort DL entre les marqueurs avantage les tests d'association allélique ;
f) si le DL entre les marqueurs à l'étude est moyen, la puissance des tests dépend de la force d'association entre la maladie et les marqueurs ;
g) une faible association avec la maladie favorise l'approche allélique ;
h) une forte association avantage les tests basés sur les haplotypes ;
i) en présence d'une association moyenne, les deux approches donnent des puissances similaires.

9.3.2 Résultats concernant les devis d'échantillonnage

Les résultats de simulation suggèrent que la puissance des tests d'association est la même dans les deux devis d'échantillonnage. Cependant, à la section 9.1.2, on a vu que les différences de puissance maximales sont généralement statistiquement significatives, en faveur du devis de cas-témoins. On en déduit que l'algorithme EM performe bien dans ce type d'échantillon puisqu'il semble inférer correctement les haplotypes de ces sujets non reliés.

Toutefois, il est important de mentionner que les différences de puissance détectées sont généralement inférieures à 5%. En pratique, des différences de puissance de cet ordre ne constituent pas un argument suffisamment convaincant pour choisir un test plutôt qu'un autre. Il faut prendre en considération d'autres facteurs tels que

- la disponibilité d'un groupe de témoins ;
- la possibilité de rencontrer un problème de stratification de la population, comme c'est souvent le cas dans des villes très cosmopolites ;
- la faisabilité du recrutement d'un échantillon de trios.

Le choix du devis doit également tenir compte de la maladie étudiée. En effet, il pourrait être difficile, voire impossible, de recruter les parents de sujets atteints d'une maladie qui survient à un âge avancé, comme la maladie d'Alzheimer par exemple. De plus, les coûts reliés au génotypage des sujets échantillonnés doivent être considérés.

L'échantillonnage de cas-témoins est avantageux par rapport à l'échantillonnage de trios. On se rappelle qu'un plus grand nombre de sujets doit être génotypé pour le devis de trios (1.5 fois plus). Supposons par exemple que le génotypage d'un marqueur coûte 0.10 $ par sujet et qu'un chercheur voulant faire une étude de parcours de génome identifie 300 000 marqueurs d'intérêt. Pour un échantillon de 100 cas et de 100 témoins, soit 200 personnes, le génotypage lui coûtera au total 6 000 000 $. En optant pour un devis d'échantillonnage de 100 trios, donc 300 personnes, il devra débourser 3 000 000 $ de plus, ce qui représente une différence non négligeable.

Tel que nous l'avons vu à la section 7.2, les scientifiques qui ont récemment étudié les tests d'association génétique dans les deux devis d'échantillonnage en sont venus à la conclusion que l'échantillon de cas-témoins est plus puissant que l'échantillon de trios. Nos résultats sont comparables. À la lumière de nos résultats, il ne semble par y avoir de différence importante entre les deux devis. Toutefois, nos échantillons de trios comportent autant de cas que les échantillons de cas-témoins, ce qui n'était pas le cas dans les travaux précédents.

9.3.3 Résultats concernant les approches analytiques

Quelle approche analytique permet de détecter l'association entre une maladie et des marqueurs avec plus de puissance ? Les résultats de notre étude de Monte-Carlo laissent entendre que la réponse dépend du niveau de DL entre les marqueurs étudiés, ainsi que de la force de l'association entre ceux-ci et la maladie.

Quand il y a un fort DL entre les marqueurs, l'information génétique est en quelque sorte redondante. Si l'on connaît l'allèle à un marqueur, on peut déduire les allèles à des marqueurs qui sont en fort DL avec celui-ci. Pour diminuer le nombre de tests ou encore le nombre de degrés de liberté des tests, on peut choisir pour les analyses un seul marqueur parmi ceux qui sont en fort DL. Les courbes de puissance des scénarios 5 et 6, pour lesquels le DL entre les marqueurs était fort, montrent d'ailleurs que les tests d'association allélique sont plus puissants que les tests d'association haplotypique dans ces conditions. Pinnaduwage et Briollais (2005) et Cordell et Clayton (2005) ont fait part de cette situation dans leurs travaux.

Si le DL entre les marqueurs est faible, on gagne à utiliser l'approche analytique basée sur les haplotypes puisque dans ce cas, l'information génétique des marqueurs n'est pas redondante. D'ailleurs, les courbes de puissance des scénarios 1 et 2, pour lesquels le DL entre les marqueurs était faible, montrent que les tests d'association haplotypique sont plus puissants que les tests d'association allélique. Les articles de Cordell et Clayton (2002) et de Pinnaduwage et Briollais (2005) mentionnent également ce fait.

Chapitre 10

Conclusion

Ce mémoire présente une étude de Monte-Carlo visant à comparer la puissance statistique de différentes stratégies d'analyse et de détection de l'association génétique basée sur le déséquilibre de liaison (DL). Des comparaisons ont été effectuées sur deux devis d'échantillonnage (cas-témoins et trios), ainsi que sur deux approches analytiques (allélique et haplotypique).

Nos simulations suggèrent que les tests visant à détecter la présence d'association entre une maladie et des marqueurs ont une puissance comparable, peu importe qu'ils soient effectués à partir d'échantillons de cas-témoins ou de trios. Pour une puissance donnée, les échantillons de cas-témoins présentent toutefois l'avantage d'être moins coûteux à génotyper, puisque le nombre de sujets sur lesquels ils s'appuient est inférieur par un facteur d'un tiers. En ce sens, les échantillons de cas-témoins sont donc plus efficaces que les trios.

Bien entendu, la supériorité des échantillons de cas-témoins n'est valable qu'à condition de pouvoir s'assurer d'une certaine homogénéité entre les cas et les témoins, particulièrement en ce qui a trait à leur origine ethnique. Ce faisant, on évite les écueils déjà évoqués au chapitre 4 concernant la stratification de la population.

Pour ce qui concerne le choix entre les tests d'association allélique et haplotypique, le niveau de DL entre les marqueurs étudiés doit être considéré. Un faible DL avantagerait les tests haplotypiques alors qu'un fort DL favoriserait les tests alléliques.

L'étude de situations plus complexes fournirait un complément utile à ces travaux. En particulier, nos conclusions sont actuellement limitées au cas de deux marqueurs à deux allèles. Cette hypothèse simpliste mais néanmoins réaliste a permis d'obtenir

des résultats dont l'interprétation s'est avérée à la fois cohérente et sans ambiguïté. Des scénarios plus élaborés pourraient toutefois être envisagés. À titre d'exemple, on pourrait considérer un plus grand nombre de marqueurs ou encore un nombre plus important d'allèles par marqueur.

Par ailleurs, les simulations effectuées dans ce mémoire supposent un mode de transmission dominant pour le gène de la maladie. En pratique, il arrive toutefois que des parents soient porteurs d'une maladie sans en être atteints. Dans des travaux ultérieurs, on pourrait donc envisager d'étendre les comparaisons entre devis et approches analytiques à des situations dans lesquelles la transmission de la maladie se fait sous un mode récessif. Plutôt que de supposer qu'un seul gène est responsable de la pathologie, on pourrait aussi considérer le cas où celle-ci résulte d'une interaction entre plusieurs gènes, phénomène connu sous le nom d'*épistasie*.

Bibliographie

Bass, M. P., Martin, E. R. et Hauser, E. R. (2004). Pedigree generation for analysis of genetic linkage and association. *Pacific Symposium on Biocomputing*, 9:93–103.

Bull, S. B., John, S. et Briollais, L. (2005). Fine mapping by linkage and association in nuclear family and case-control designs. *Genetic Epidemiology*, 29 Suppl. 1:S48–S58.

Chromosome 3q Registry (2006). Family information and support group. [En ligne], http://members.cox.net/chromosome3/ (Page consultée le 25 juillet 2006).

Clayton, D. G. et Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

Cordell, H. J. et Clayton, D. G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data : Application to HLA in type 1 diabetes. *American Journal of Human Genetics*, 70:124–141.

Cordell, H. J. et Clayton, D. G. (2005). Genetic epidemiology 3 : Genetic association studies. *The Lancet*, 366:1121–1131.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A., Haines, J. et Pericak-Vance, M. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261:921–923.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38.

Dudbridge, F., Koeleman, B. P., Todd, J. A. et Clayton, D. G. (2000). Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *American Journal of Human Genetics*, 66:2009–2012.

Excoffier, L. et Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12:921–927.

Falk, C. et Rubenstein, P. (1987). Haplotype relative risks : An easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics*, 51:227–233.

FBAT (2006). A program for family-based association testing. [En ligne], http://www.biostat.harvard.edu/~fbat/fbat.htm (Page consultée le 25 juillet 2006).

Funke, B., Finn, C. T., Plocik, A. M., Lake, S., DeRosse, P., Kane, J. M., Kucherlapati, R. et Malhotra, A. K. (2004). Association of the DTNBP1 locus with schizophrenia in a U.S. population. *American Journal of Human Genetics*, 75:891–898.

Genome News Network (2006). Genome variations. [En ligne], http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp4_1.shtml (Page consultée le 25 juillet 2006).

Gloyn, A. L., Weedon, M. N., Owen, K. R., Turner, M. J., Knight, B. A., Hitman, G., Walker, M., Levy, J. C., Sampson, M., Halford, S., McCarthy, M. I., Hattersley, A. T. et Frayling, T. M. (2003). Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes*, 52:568–572.

Guo, S. W. et Thompson, E. A. (1992). Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, 48:361–372.

Hintsanen, P., Sevon, P., Onkamo, P., Eronen, L. et Toivonen, H. (2006). An empirical comparison of case-control and trio-based study designs in high-throughput association mapping. *Journal of Medical Genetics*, 43:617–624.

Horvath, S., Xu, X. et Laird, N. M. (2000). The family based association test method : Computing means and variances for general statistics. Rapport technique, Department of Biostatistics, Harvard School of Public Health, Boston, MA.

Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J. P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J.-F., Sahbatou, M. et Thomas, G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411:599–603.

Human Genome Project (2006). [En ligne], http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml (Page consultée le 25 juillet 2006).

International HapMap Project (2006). Le projet international hapmap. [En ligne], http://www.hapmap.org/ (Page consultée le 25 juillet 2006).

Laird, N. M., Horvath, S. et Xu, X. (2000). Implementing a unified approach to family based tests of association. *Genetic Epidemiology*, 19 Suppl. 1:S36–S42.

Lampron, P. (2006). Site pédagogique visant une introduction à la génétique. [En ligne], http://www3.sympatico.ca/philipe.lampron/ (Page consultée le 25 juillet 2006).

Lewontin, R. (1964). The interaction of selection and linkage. I. General considerations ; heterotic models. *Genetics*, 49:49–67.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M. et Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266:66–71.

Pinnaduwage, D. et Briollais, L. (2005). Comparison of genotype- and haplotype-based approaches for fine-mapping of alcohol dependence using COGA data. *BMC Genetics*, Suppl. 6:65.

Risch, N. et Merikangas, K. R. (1996). The future of genetic studies of complex human diseases. *Science*, 273:1516–1517.

Schaid, D. J. (1996). General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology*, 13:423–449.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. et Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics*, 70:425–434.

Shephard, N., John, S., Cardon, L., McCarthy, M. I. et Zeggini, E. (2005). Will the real disease gene please stand up ? *BMC Genetics*, Suppl. 6:66.

Spielman, R. S., McGinnis, R. E. et Ewens, W. J. (1993). Transmission test for linkage disequilibrium : The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, 52:506–516.

Stefansson, H., Sarginson, J., Kong, A., Yates, P., Steinthorsdottir, V., Gudfinnsson, E., Gunnarsdottir, S., Walker, N., Petursson, H., Crombie, C., Ingason, A., Gulcher, J. R., Stefansson, K. et St Clair, D. (2003). Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. *American Journal of Human Genetics*, 72:83–87.

The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72:971–983.

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

Weir, B. S. (1990). *Genetic Data Analysis*. Sinauer, Sunderland, MA.

Wilson, R. D. (2002). Le dépistage des porteuses de fibrose kystique durant la grossesse au Canada. *Journal d'obstétrique et gynécologie du Canada*, 24:648–651.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C. et Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378:789–792.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J. et Ehm, M. G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*, 53:79–91.

Zhao, J. H., Curtis, D. et Sham, P. C. (2000). Model-free analysis and permutation tests for allelic associations. *Human Heredity*, 50:133–139.