Laboratoire
Statistique et Génome

2007EVRY0015

Thèse CIFRE présentée pour l'obtention du titre de
Docteur de l'Université d'Évry-Val d'Éssone
en **Statistique Génétique** par

Mickaël Guedj

Méthodes Statistiques pour l'Analyse des Données Génétiques d'Association à Grande Échelle

Soutenue le 13 Juillet 2007 devant le jury composé de :

Directeurs de thèse :	Grégory Nuel	Maître de Conférence en Statistique à l'Université d'Évry-Val D'Éssone
	Bernard Prum	Professeur de Statistique à l'Université d'Évry-Val D'Éssone
Encadrant Serono :	Jérôme Wojcik	Responsable de la Bioinformatique chez Serono, Génève
Présidente du jury :	Florence Demenais	Directrice d'unité INSERM, Evry
Rapporteurs :	David Balding	Professeur de Statistique à l'*Imperial College* de Londres
	Avner Bar-Hen	Professeur de Statistique à l'Université de Paris XIII
Examinateurs :	Françoise Clerget	Directrice d'unité INSERM, Villejuif
	Jean-François Zagury	Professeur de Bioinformatique au Centre National des Arts et Métiers

Remerciements

Je tiens à remercier en premier lieu Grégory Nuel pour avoir dirigé ce travail de thèse en y apportant ses compétences en Statistique et Algorithmique ainsi que son point de vue sur la Génétique. Travailler sous sa direction a été très enrichissant, tant par ses qualités d'enseignant que de chercheur, ainsi que pour l'énergie et l'engouement dont il sait faire preuve.

Je remercie également Bernard Prum pour sa pédagogie, sa gentillesse, et pour les responsabilités qu'il m'a confiées, me permettant ainsi de prendre pleinement part à l'essor de la thématique Statistique Génétique au sein de son équipe.

Je voudrais remercier Jérôme Wojcik pour son encadrement au sein de *Serono*, ses suggestions, ainsi que pour l'intérêt qu'il a continuellement manifesté envers mon travail en l'intégrant aux projets de recherche de *Serono*.

Je remercie également Hiroaki Tanaka pour avoir largement contribué à initier l'échange scientifique entre *Serono* et le laboratoire *Statistique et Génome* et donc permis la réalisation de cette thèse à l'interface de ces deux entités de recherche.

De manière plus générale, je suis très reconnaissant envers l'ensemble de ces responsables pour le temps et la liberté de travail qu'ils ont su m'accorder.

Je remercie les membres du jury pour m'avoir fait l'honneur de leur présence à ma soutenance et m'avoir permis, par leurs conseils et leurs remarques, de terminer la rédaction de ce manuscrit.

Je remercie les personnes avec qui j'ai eu l'occasion de discuter ou de collaborer et qui ont ainsi largement contribué à ce travail. En particulier je voudrais remercier K Forner, D Robelin, F Picard, E Della-Chiesa, M Hoebeke, M Lamarine, H Aschard, J Dauvillier, V Mièle, C Ambroise, S Robin, S Lèbre, C Matias, A Guilbot, C Stalens, Y Slaoui, B Junot, M Baudry et M Ilbert ainsi que l'ensemble du personnel du laboratoire *Statistique et Génome* et de *Serono*.

Je suis très reconnaissant envers le CNRS, l'INRA, l'Université d'Evry-Val d'Essonne, *Serono* et l'ANRT pour avoir financé ma thèse ainsi que les nombreux déplacements qu'elle a suscités. Je voudrais remercier une dernière fois les responsables de *Serono* ainsi que les

membres du comité GAW-15[1] pour m'avoir permis de disposer librement de leur données.

Enfin, pour leur soutien non-scientifique mais non moins significatif, je tiens à remercier mes parents, mon frère, la famille Pecnik, Natassia, Arielle, Claire, Émilie, Johanne, William et les BIMs.

[1]en la personne de Michael Miller et du support apporté par NIH *grants* 5RO1-HL049609-14, 1201-AG021917-01A1, l'université du Minnesota, le *Minnesota Supercomputing Institute* et le GAW *grant* R01-GM031575.

Résumé

Les avancées en Biologie Moléculaire ont accéléré le développement de techniques de génotypage haut-débit et ainsi permis le lancement des premières études génétiques d'association à grande échelle. La dimension et la complexité des données issues de ce nouveau type d'étude posent aujourd'hui de nouvelles perspectives statistiques et informatiques nécessaires à leur analyse, constituant le principal axe de recherche de cette thèse.

Après une description introductive des principales problématiques liées aux études d'association à grande échelle, nous abordons plus particulièrement les approches simple-marqueur avec une étude de puissance des principaux tests d'association, ainsi que de leur combinaisons. Nous considérons ensuite l'utilisation d'approches multi-marqueurs avec le développement d'une méthode d'analyse fondée à partir de la statistique du Score Local. Celle-ci permet d'identifier des associations statistiques à partir de régions génomiques complètes, et non plus des marqueurs pris individuellement. Il s'agit d'une méthode simple, rapide et flexible pour laquelle nous évaluons les performances sur des données d'association à grande échelle simulées et réelles. Enfin ce travail traite également du problème du test-multiple, lié au nombre de tests à réaliser lors de l'analyse de données génétiques ou génomiques haut-débit. La méthode que nous proposons à partir du Score Local prend en compte ce problème. Nous évoquons par ailleurs l'estimation du *Local False Discovery Rate* à travers un simple modèle de mélange gaussien.

L'ensemble des méthodes décrites dans ce manuscrit ont été implémentées à travers trois logiciels disponibles sur le site du laboratoire *Statistique et Génome* : `fueatest`, `LHiSA` et `kerfdr`.

Abstract

The increasing availability of dense Single Nucleotide Polymorphisms (SNPs) maps due to rapid improvements in Molecular Biology and genotyping technologies have recently led geneticists towards genome-wide association studies with hopes of encouraging results concerning our understanding of the genetic basis of complex diseases. The analysis of such high-throughput data implies today new statistical and computational problematics to face, which constitute the main topic of this thesis.

After a brief description of the main questions raised by genome-wide association studies, we deal with single-marker approaches by a power study of the main association tests and their combination. We consider then the use of multi-markers approaches by focusing on the method we developed which relies on the Local Score. This sum statistic identifies associations between regions and the disease instead of marker considered individually. It represents a simple, fast and flexible method for which we assess the efficiency based on simulated and real genome-wide association data. Finally, this thesis also deals with the multiple-testing problem attached to the number of independent tests performed for the analysis of high-throughput data. Our Local Score-based approach circumvents this problem by reducing the number of tests. In parallel, we present an estimation of the Local False Discovery Rate by a simple Gaussian mixed model.

The methods described in this manuscript are implemented in three softwares available on the website of the *Statistique et Génome* laboratory : `fueatest`, `LHiSA` and `kerfdr`.

Table des matières

Remerciements **i**

Résumé **iii**

Abstract **v**

Préambule **1**
Contexte et objectifs 1
Plan de la thèse 3

1 Introduction **5**
1.1 Préceptes de Statistique : le test d'hypothèse 6
1.2 Préceptes de Génétique 7
1.3 Épidémiologie Génétique 14
1.4 Cadres d'étude 16
1.5 Déroulement d'une étude d'association *genome-wide* 22
1.6 Contrôle qualité : validité et fiabilité des résultats 28

2 Approches simple-marqueur **39**
2.1 Introduction 40
2.2 Association statistique et tests d'indépendance 41
2.3 Tests d'association marqueur-maladie 51
2.4 Étude de puissance 56
2.5 Cas particulier du test allélique 69
2.6 Cas particulier du test d'Hardy-Weinberg 77
2.7 FDR Local 87
2.8 Conclusions 99

3 Approches multi-marqueurs **103**
3.1 Introduction 104
3.2 Approches multi-locus existantes 106
3.3 Score Local 114
3.4 Algorithme LHiSA 122
3.5 Applications 124
3.6 Discussion sur le Score Local 132
3.7 Conclusions 134

4 Conclusions **137**
4.1 Conclusions générales 138
4.2 Perspectives 142
4.3 De l'Épidémiologie Génétique à l'Épidémiologie Génomique 144

Communication scientifique **147**

Annexes **189**

Bibliographie **200**

Préambule

Contexte et objectifs

Les industries pharmaceutiques ont toujours eu pour objectif de développer des méthodes permettant la production de nouvelles molécules thérapeutiques. Depuis longtemps et encore aujourd'hui, l'approche pharmacologique privilégiée consiste à choisir un grand nombre de molécules en suivant un cahier des charges, de les tester et d'en observer les effets. On parle alors de criblage ou *screening*. Si celui-ci a dans le passé largement fait ses preuves, il commence à montrer ses limites face à de nouvelles approches s'appuyant sur la compréhension des mécanismes biologiques à l'origine des maladies.

En effet, on sait aujourd'hui qu'un grand nombre de pathologies ont une composante génétique. Une séquence d'ADN pouvant être représentée comme une succession de lettres {a,t,g et c} appelées bases, en une même position d'un chromosome ou locus, on pourra trouver différentes versions du "texte génétique" ou allèles. Dans le cas le plus simple, le changement d'un allèle en un locus donné peut à lui seul être responsable d'une maladie : on parle alors de maladie monogénique ; de manière moins triviale, la maladie dite multi-factorielle ou complexe est le résultat de composantes multi-géniques et environnementales, ce qui est le cas de la plupart des cancers, des maladies psychiatriques, auto-immunes et bien d'autres. Les études épidémiologiques fondées sur la Génétique cherchent donc à localiser les locus de susceptibilité et à en évaluer la responsabilité. Une approche populaire consiste à collecter un échantillon d'individus appelés "cas" et d'individus non-affectés appelés "témoins" et de déterminer les positions dans le génome pour lesquelles le texte génétique diffère significativement entre les cas et les témoins. On parle alors d'étude d'association cas-témoins. Il n'est bien entendu pas question aujourd'hui de séquencer le génome entier de tout individu introduit dans l'étude, bien que la faisabilité d'une telle démarche soit de plus en plus concevable. Il s'agit plutôt de travailler à partir d'un jeu de locus appelés marqueurs génétiques, dont la position sur le génome est connue et la configuration génotypique est techniquement "facile" à déterminer.

Les *Single Nucleotide Polymorphisms* communément appelés SNPs, sont des marqueurs constituant une source d'information génétique riche et abondante. Définis comme des positions sur le chromosome où le texte génétique varie d'une seule base d'un individu

à un autre, ils sont nombreux, apparaissant en moyenne une fois toutes les 2,000 bases le long des 3 milliards de bases qui constituent le génome humain. La diminution à la fois du coût et du temps de génotypage des SNPs a contribué récemment au lancement d'études génétiques d'association à grande échelle permettant d'explorer une part conséquente des polymorphismes génétiques pouvant être impliqués dans les mécanismes biologiques à l'origine des maladies. L'analyse de telles données n'est pas sans soulever de nombreuses questions méthodologiques.

A la suite de résultats encourageants sur la schizophrénie mettant en cause les deux gènes *G72* et *DAAO* (Chumakov et al 2002), l'industriel *Serono*, par l'intermédiaire de son *Serono Genetics Institute* localisé à Évry, se lance dans une vaste étude d'association cas-témoins concernant quatre maladies auto-immunes : la sclérose en plaque, la polyarthrite rhumatoïde, le lupus et le psoriasis. Le but est à la fois de comprendre l'étiologie de chaque maladie afin de déterminer de nouvelles cibles thérapeutiques, mais aussi de mettre en avant des mécanismes communs aux quatre maladies. Ce projet porte le nom de projet AIM-Scan.

A l'origine de *Serono Genetics Institute* on trouve l'entreprise française de biotechnologies *Genset*, fondé en 1989 et l'une des premières grandes entreprises présentes sur le site de la génopole d'Evry. En 2002, *Genset* rejoint le groupe *Serono* en devient alors le *Serono Genetics Institute*, lequel compte 130 employés dont 90% sont rattachés à la recherche. En 2006, *Serono* ferme le site d'Évry et transfert l'ensemble de ses activités à Genève. Début 2007, *Serono* fusionne avec l'industriel pharmaceutique *Merck* pour former le groupe *Merck-Serono*.

Pour le projet AIM-Scan, *Serono* passe un accord avec *Affymetrix* afin d'utiliser avec un temps d'avance les puces de génotypages 100K et 500K développées par l'industriel. Pour chaque maladie, *Serono* dispose d'échantillons issus de deux à trois populations indépendantes et dont la taille peut aller jusqu'à 1,000 individus cas et témoins. Avec un tel jeu de données en main et voulant se donner toutes les chances de réussir, *Serono Genetics Institute* se tourne en 2004 vers le laboratoire *Statistique et Génome* afin de mettre en commun leurs expertises respectives des études d'association pour l'un, et de l'analyse statistique des données génomiques pour l'autre. C'est dans ce contexte que se situe cette thèse CIFRE entre l'industriel et le laboratoire.

Les objectifs de cette thèse, tels qu'ils ont été définis à l'origine, ont été d'apporter un support en Statistique ainsi qu'une connaissance de la littérature dans le but d'améliorer la compréhension des méthodes d'analyse des études d'association cas-témoins existantes. Ils ont également été de contribuer au passage à l'analyse de données à grande échelle en y apportant un développement méthodologique adapté, répondant aux besoins identifiés par *Serono* et essentiellement guidé par les exigences soulevées par l'analyse de telles données.

L'analyse de données d'association à grande échelle représente une nouvelle thématique de recherche pour le laboratoire *Statistique et Génome*. Celle-ci a débuté lors de la collaboration avec *Serono* et s'est concrétisée avec la mise en place de cette thèse CIFRE.

Plan de la thèse

Ce manuscrit s'organise principalement autour de quatre grands chapitres. Chaque chapitre s'ouvre sur un énoncé introductif des différents points qui y sont abordés et se conclut sur une synthèse. Le travail présenté est par ailleurs accompagné d'un développement logiciel conséquent, avec la mise à disposition des méthodes proposées à la communauté. Dans cette esprit, nous traitons à plusieurs reprises de leur mise en oeuvre pratique ; cet aspect nous semble en effet essentiel lorsqu'on est confronté à l'analyse de données mettant en jeu de sérieuses exigences en terme de complexité algorithmique. Enfin cette thèse s'appuie principalement sur les données apportées par *Serono*. Pour des raisons de confidentialité évidentes, l'utilisation des données garde une dimension uniquement illustrative et nous n'indiquons en conséquence aucun résultats "biologiques" obtenus avec nos méthodes. Nous sommes néanmoins très reconnaissants envers les responsables de *Serono* de nous avoir permis d'utiliser ces données avec autant de liberté.

Afin de faciliter la lecture de ce manuscrit et de la rendre plus fluide, les points les plus techniques de Statistique sont ramenés en annexe.

Le chapitre 1 est un chapitre d'introduction. Il permet de poser les fondements statistiques et génétiques nécessaires à la compréhension de la démarche scientifique sur laquelle s'appuie toute étude en Épidémiologie Génétique. Nous présentons à cette occasion les différents cadres d'études possibles en insistant sur les principaux. En particulier l'accent est mis sur les études d'association cas-témoins à grande échelle, dont l'analyse constitue la problématique principale de cette thèse. Nous déroulons également les grandes étapes d'une étude : la génération des données, l'analyse statistique, la formulation d'hypothèses et leur confirmation par réplication. Enfin nous introduisons et discutons les principaux facteurs évoqués dans la littérature qui peuvent affecter la qualités des résultats : le manque de puissance, le test-multiple, les erreurs liées au génotypage, les valeurs manquantes dans les données ainsi que la stratification de population. Ce chapitre se termine sur le défi que pose parallèlement toute la complexité intrinsèque aux données auxquelles on s'intéresse.

La suite du manuscrit détaille plus spécifiquement le travail de recherche réalisé à l'occasion de cette thèse inspiré des problématiques méthodologiques soulevées par *Serono* pour l'analyse de leurs données d'association cas-témoins *genome-wide*.

Le chapitre 2 réunit l'ensemble du travail réalisé sur les approches dites simple-marqueur qui traitent chaque marqueur individuellement. Il existe un certain nombre de tests d'association et donc de stratégies d'analyse possibles. A travers une étude de puissance nous discutons leur pertinence. En particulier l'accent est mis sur la comparaison des différents modes d'estimation de la puissance ainsi que sur la validité statistique des tests allélique et d'Hardy-Weinberg pour lesquels on propose plusieurs alternatives. Ensuite nous nous intéressons au problème du test-multiple en présentant une quantité statistique introduite récemment, et qui nous semble apporter une information intéres-

sante dans le cadre des études d'association *genome-wide* : le *Local False Discovery Rate*. Nous abordons son estimation à travers un modèle de mélange gaussien qui constitue pour nous la méthode la plus simple et la plus intuitive de considérer le problème. Ce chapitre traitant d'aspects assez variés des approches simple-marqueur, il s'achève sur une synthèse qui permet de les lier et de les replacer dans le contexte actuel des études d'association.

Le chapitre 3 traite du problème de l'analyse simultanée de plusieurs marqueurs, des principaux enjeux méthodologiques soulevés par ce type d'approches et des différentes solutions proposées dans la littérature. Il décrit également l'élément qui constitue pour nous le développement méthodologique le plus important réalisé durant cette thèse : une méthode d'analyse multi-marqueurs construite à partir de la statistique du Score Local. S'appuyant sur la détection d'accumulations de signaux d'association élevés autour de locus de susceptibilité, cette approche permet d'effectuer sur un grand nombre de marqueurs ce que l'expert a tendance à réaliser "à l'oeil" sur de plus petits jeux de données. Cette nouvelle méthode est appliquée sur quatre jeux de données réels et simulés. Enfin, sur la base des résultats obtenus, elle est discutée et replacée dans le contexte défini par les approches multi-marqueurs existantes.

Le dernier chapitre de conclusion a plusieurs objectifs : il reprend naturellement l'ensemble des points développés dans cette thèse mais ouvre également sur différentes perspectives scientifiques quant à la l'analyse des données d'association *genome-wide*. Ce chapitre achève ce manuscrit sur une discussion concernant l'apport des études d'association à grande échelle en Épidémiologie Génétique et comment elles s'intègrent dans une démarche plus générale d'acquisition de connaissances permettant d'élucider les mécanismes biologiques à l'origine des maladies complexes.

Chapitre 1

Introduction

L'objectif de ce chapitre introductif est de poser les bases nécessaires à la compréhension du travail de recherche réalisé dans le cadre de cette thèse. En particulier il permet de poser le contexte scientifique dans lequel se situent aujourd'hui les études génétiques d'association *genome-wide*.

Dans un premier temps nous introduisons un certain nombre de préceptes statistiques et génétiques fondamentaux tels que le test d'hypothèse, la diversité génétique, le déséquilibre de liaison et l'équilibre d'Hardy-Weinberg. Nous présentons ensuite l'ensemble des thématiques scientifiques que recouvre l'Épidémiologie Génétique, ainsi que les différents cadres d'étude possibles : familiale/cas-témoins, liaison/association, gènes-candidats/*genome-wide*. Puis, nous évoquons chacune des étapes d'une étude d'association : **(i)** le recrutement des individus (à l'occasion duquel nous abordons également des questions d'ordre éthique), **(ii)** le choix des marqueurs à génotyper ainsi que les techniques modernes de génotypage, **(iii)** l'analyse statistique, la formulation d'hypothèses et **(iv)** la vérification des hypothèses énoncées par réplication.

Prolongeant la constatation que la réplication de résultats est en pratique difficile à obtenir, la dernière section de cette introduction traite du contrôle qualité d'une étude, c'est à dire de tous les facteurs pouvant affecter la validité et la fiabilité des résultats. En particulier le manque de puissance, le test-multiple, les erreurs de génotypages, les valeurs manquantes et la stratification de la population, ont été souvent mis en avant dans la littérature. Au delà de ces facteurs relatifs au *design* de l'étude ou à la qualité des données, la complexité des étiologies pose également un défi pour l'élucidation des mécanismes biologiques mis en cause.

Notes bibliographiques : la rédaction de ce chapitre s'est appuyée en partie sur la lecture de Garnier (2007), Balding (2006), Newton-Cheh et Hirschhorn (2005), Hirschhorn et Daly (2005), Shen et al (2005), Jannot (2004), Sillanpaa et Auranen (2004), Page et al (2003) et Elston et al (2002).

1.1 Préceptes de Statistique : le test d'hypothèse

Définition

Une question essentielle dans une démarche scientifique est souvent d'établir une relation entre deux concepts, qu'il s'agisse d'une association ou d'une comparaison. Cette démarche passe dans un premier temps par l'élaboration d'hypothèses, puis par leur validation.

Une façon assez naturelle bien que fausse de raisonner est de se dire : *A implique B donc B implique A*. En réalité ce raisonnement est faux dans la mesure où une alternative A' permettrait aussi de justifier l'observation de B. Une démarche de test d'hypothèse adoptera donc plutôt une stratégie de démonstration par l'absurde en cherchant à montrer la fausseté de B pour en déduire que A n'est pas vrai : *non B implique non A*. Cette dernière hypothèse, qui réfère à la négation de l'hypothèse de recherche ($H1$), est l'hypothèse nulle ($H0$).

Prise de décision

Compte tenu d'une hypothèse nulle, quatre situations sont possibles quant à l'issue du test de cette hypothèse : on peut décider de rejeter ou non $H0$, alors qu'en réalité (mais nous ne le savons pas) $H0$ est vraie ou fausse. On se trouve donc dans l'incertitude quant à la décision à prendre, et l'enjeu sera alors de se convaincre que l'on prend la bonne décision en contrôlant le risque de se tromper.

Il existe en réalité deux façons distinctes de se tromper et de fait, deux types de risques. On peut rejeter $H0$ alors que $H0$ est vraie, c'est à dire affirmer une association ou une différence alors qu'il n'y a rien ; ce type d'erreur est appelé **erreur de type-I** et le risque associé est le taux d'erreur de type-I noté α[1]. Si d'autre part on décide de ne pas rejeter $H0$ alors que $H1$ est vraie, je commets alors une **erreur de type-II** avec un taux noté β[2].

	H_0 non rejetée	H_0 rejetée
H_0 vraie	$1-\alpha$	α (erreur de type-I)
H_0 fausse	β (erreur de type-II)	$1-\beta$

En pratique, la valeur de β dépend de l'alternative $H1$ et il est quasiment impossible de l'estimer en toute généralité. C'est pourquoi, seul α est utilisé comme critère de décision.

[1] on parle également d'erreur et de risque de première espèce
[2] on parle également d'erreur et de risque de deuxième espèce

Par ailleurs, $1-\beta$ s'appelle la puissance d'un test. On comparera - quand cela est possible - deux tests en comparant leur puissance.

Dans tous les cas, un test d'hypothèse suit un succession d'étapes définies : **(i)** énoncé des hypothèses nulle et alternative ; **(ii)** calcul d'une variable de décision - la **statistique** ($\mathcal{S}$) - correspondant à une fonction des observations. Elle mesure une distance entre ce que j'observe et ce que j'attends sous l'hypothèse nulle. Plus cette distance est grande et moins $H0$ est probable ; **(iii)** calcul de la **probabilité critique** (ou p-value[3]) d'obtenir une valeur observée de la statistique ($\mathcal{S}^{\text{obs}}$) au moins aussi élevée que la valeur obtenue si $H0$ est vraie :

$$p\text{v} = \mathbb{P}_{H0}(\mathcal{S} \geqslant \mathcal{S}^{\text{obs}});$$

(iv) conclusion du test en fonction de la valeur de la p-value par rapport à une valeur seuil du risque de première espèce (α) ou **niveau** du test. La conclusion peut se faire de façon analogue sur la base de la statistique elle-même par rapport à un seuil t_α tel que :

$$\alpha = \mathbb{P}_{H0}(\mathcal{S} \geqslant t_\alpha).$$

Types de tests

La pratique des tests statistiques nécessite que l'on distingue différentes situations. Celles-ci sont décrites par trois éléments : **(i)** la forme du test (comparaison bilatérale ou unilatérale), **(ii)** la possibilité de faire appel à une loi de distribution connue (test paramétrique ou non-paramétrique) et **(iii)** l'appariement des mesures (une ou plusieurs mesures réalisées sur un même échantillon).

1.2 Préceptes de Génétique

Le génome, siège de l'information génétique

La Génétique est la science qui étudie la transmission des caractères des parents à leurs enfants. Depuis la fin du XIXème siècle, les mécanismes de l'hérédité sont de mieux en mieux compris : chaque individu porte en chacune de ses cellules un patrimoine génétique qui détermine un grand nombre de ses caractéristiques. Ce patrimoine, qui est appelé le **génome**, est composé d'une ou plusieurs entités appelées les **chromosomes**. Le nombre de chromosomes dépend de l'espèce ; les bactéries n'ont par exemple qu'un seul chromosome tandis que l'espèce humaine en compte 46 : 22 paires de chromosomes homologues et 2 chromosomes sexuels.

[3] notée pv

Chacun de ces chromosomes est en fait une chaîne orientée le long de laquelle se succèdent quatre molécules différentes appelées bases ou nucléotides et notées A, C, G et T pour adénine, cytosine, guanine et thymine. On parle alors de **séquence d'ADN**[4]. Un chromosome est donc un texte écrit dans l'alphabet constitué de ces quatre lettres. Une part de ce génome permet, selon un code aujourd'hui parfaitement déchiffré, la fabrication par la cellule de molécules participant à tous les mécanismes du vivant (la respiration, l'alimentation...) : les **protéines**. Ce qui code pour une protéine est appelé un **gène**. Mais seule une proportion limitée du génome code pour les protéines et un même gène peut coder pour plusieurs protéines. Depuis que l'on connaît la séquence complète du génome humain (*Human Genome Project*[5]) on estime entre environ 20,000 et 25,000 le nombre de gènes présents chez l'Homme, ce qui représente à peu près 5% du génome. Les 95% restants contiennent des éléments de régulation de l'expression génique ainsi qu'une grande quantité d'ADN dont la fonction reste à déterminer.

Diversité génétique

On désigne par **locus** une position du génome et par **allèle** une version donnée du texte génétique. Un **polymorphisme** correspond alors à la présence en un locus de plusieurs allèles. On définit par haplotype la combinaison de plusieurs allèles situés sur des locus différents d'un même chromosome. Dans l'espèce humaine, chaque individu possède 22 paires de chromosomes homologues ; on trouvera donc en un locus donné une combinaison de deux allèles que l'on appelle **génotype**. Ces 44 chromosomes auxquels s'ajoutent 2 chromosomes sexuels, totalisent environ 3 milliards de bases et de fait, seule une infime partie du génome varie d'un individu à l'autre. La diversité génétique au sein d'une population est essentiellement due à deux événements : la mutation et la recombinaison.

- Mutation : l'introduction de polymorphismes dans le patrimoine génétique d'une population est le résultat d'événements de mutation. Une mutation correspond à une modification soudaine et transmissible de la séquence d'ADN, par exemple le changement, l'ajout/insertion ou la suppression/délétion d'une base. En fonction de la base affectée, ces mutations peuvent être silencieuses, c'est à dire n'avoir aucun effet sur la protéine résultante, ou au contraire avoir une incidence positive ou négative sur la protéine et donc sur l'individu. Par exemple la mutation d'un gène peut changer la constitution, la forme et par la même occasion la fonction de la protéine correspondante ; on parle alors de mutation *missense*.

- Recombinaison : une autre source de diversité génétique est la recombinaison. Il s'agit d'un phénomène qui se produit par enjambement des chromosomes homologues[6].

[4] Acide Désoxyribo-Nucléique
[5] `http://www.sanger.ac.uk/HGP`
[6] *crossing-over*

Elle survient au cours du processus de formation des gamètes[7] : la méiose. Chaque chromosome a alors la possibilité d'échanger une partie d'ADN avec son chromosome homologue. La chance qu'un événement de recombinaison se produise entre deux locus augmente avec la distance qui les sépare.

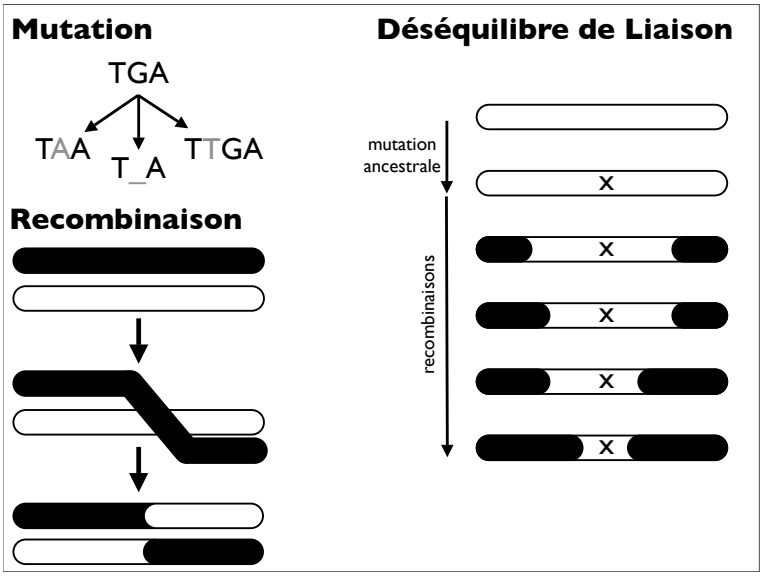


Fig. 1.1 – **Mutations :** elle peut se traduire par un changement de base, une délétion ou une insertion. **Recombinaison :** elle est provoquée par l'enjambement de chromosomes homologues au cours de la production des gamètes. **Déséquilibre de liaison :** il est le résultat d'un événement de mutation ancestral et d'une succession de recombinaisons au cours des générations.

Déséquilibre de Liaison

- Définition : le déséquilibre de liaison noté LD pour *Linkage Disequilibrium* décrit la relation entre deux allèles à deux locus dans une population et existe lorsque la probabilité d'observer un couple d'allèles sur un chromosome n'est pas égale au produit des probabilités d'observer ces allèles individuellement. Il se définit également entre plus de deux locus. Le LD peut avoir différentes origines. Le plus fréquemment il survient lorsqu'un nouvel allèle apparaît par mutation dans une région chromosomique caractérisée, dans la population, par un *pattern* d'allèles ou **haplotype** spécifique sur les locus voisins. Ce *pattern* d'allèles nouvellement constitué est transmis en bloc de générations en générations, plus ou moins altéré par les recombinaisons successives. Par conséquent, les locus à proximité du locus ayant initialement muté conservent alors une forte **association allélique** qui caractérise le LD présent dans cette portion du génome.

[7] cellules impliquées dans la reproduction

- Mesures de LD : les mesures de LD sont en fait des mesures d'association (voir chapitre 2 p. 39) qui vont quantifier l'écart entre les proportions haplotypiques observées et celles attendues sous l'hypothèse d'indépendance entre les allèles. Il en existe un certain nombre ce qui ne facilite pas les comparaisons entre différentes études. Nous introduisons ici les trois principales.

Soit deux locus bi-alléliques dont les allèles sont a/A et b/B respectivement. Soit p_a, p_A, p_b et p_B les proportions alléliques et p_{ab}, p_{aB}, p_{Ab} et p_{AB} les proportions haplotypiques correspondantes. Le coefficient de déséquilibre de liaison ($\mathcal{D}$) correspond à la simple différence entre la proportion d'un haplotype donné et celle attendue sous l'hypothèse d'indépendance :

$$\mathcal{D} = p_{AB} - p_A p_B = p_{ab} - p_a p_b.$$

Ainsi plus $\mathcal{D}$ est élevé et plus les locus sont en déséquilibre de liaison. Des standardisations de $\mathcal{D}$ ont été proposées afin d'avoir des mesures comprises entre -1 et 1. Les plus connues sont le coefficient $\mathcal{D}'$ de Lewontin (1964) et le coefficient de corrélation r^2.

Le déséquilibre de liaison est dépendant à la fois du temps écoulé depuis la mutation initiale et du taux de recombinaison entre les deux locus. $\mathcal{D}'$ permet d'estimer de combien le déséquilibre a diminué par rapport à sa valeur initiale :

$$\mathcal{D}' = \frac{\mathcal{D}}{\mathcal{D}_{\max}} \text{ avec } \mathcal{D}_{\max} = \begin{cases} \min(p_A p_b; p_a p_B) & \text{si } D > 0 \\ \min(p_a p_b; p_A p_B) & \text{si } D < 0 \end{cases}$$

$\mathcal{D}'$ a la propriété de prendre les valeurs 1 ou -1 lorsque les deux allèles n'ont pas été séparés au cours de l'histoire de la population. En pratique, cela se traduit par l'absence d'un des haplotypes possibles ; on parle alors de **déséquilibre complet**. Prenons un exemple :

	b	B
a	0.1029	0.0719
A	0.8252	0

Pour cet exemple précis, $\mathcal{D} = -0.0593$ et $\mathcal{D}' = -1$. Si des valeurs inférieures à 1 indiquent intuitivement que l'association allélique initiale a été dégradée, elle n'ont cependant pas d'interprétation précise et dépendent directement de la taille de l'échantillon, ce qui rend impossible toute comparaison entre différentes études. Par ailleurs une valeur de 1 pour $\mathcal{D}$ ou $\mathcal{D}'$ n'implique pas que deux locus portent exactement la même information. Pour cette raison, on utilise aujourd'hui plutôt un indice de corrélation (r^2), lié à la quantité d'information que fournit un locus sur l'autre :

$$r^2 = \frac{\mathcal{D}^2}{p_a p_A p_b p_B}.$$

Une valeur de 1 ne peut être observée que si l'information portée par un marqueur apporte une idée complète de celle portée par le second. En pratique cela se traduit par la présence de seulement deux des génotypes possibles et par l'égalité des proportions alléliques ($p_a = p_b$). On parle alors de **déséquilibre parfait**. Prenons un exemple :

	b	B
a	0.5763	0
A	0	0.4237

Ici, $\mathcal{D} = 0.2442$, $\mathcal{D}' = 1$ et $r^2 = 1$ alors que sur l'exemple précédent (déséquilibre complet) r^2 valait 0.6044.

En pratique, on connaît les génotypes pour chaque individu sans indication de phase, c'est à dire sans la connaissance du chromosome sur lequel se trouve chacun des deux allèles. Les proportions haplotypiques observées dans un échantillon ne sont donc pas connues, ce qui pose un problème pour l'estimation du déséquilibre de liaison. Certaines méthodes d'estimation de données incomplètes permettent cependant de les estimer à partir des proportions génotypiques (Excoffier et Slatkin 1995, Stephens et al 2001, Coulonges et al 2006).

- Blocs de LD : A première vue, on peut penser que le déséquilibre de liaison décroît avec la distance, manifestation des événements de recombinaison ayant eu lieu au cours de l'histoire de la population (Collins et al 2001, figure 1.2-A p. 13). Certaines études fondées sur des simulations montrent qu'un déséquilibre significatif ne s'étendait généralement pas au delà de 3kb (Goldstein 2001). Cependant en pratique, de fortes valeurs de LD sont observées au delà de 500kb. Il n'existe donc pas vraiment de logique en ce qui concerne le degré de LD entre deux marqueurs plus ou moins distants.

Des études plus récentes montrent une structuration du génome en fonction du LD résultant d'un taux de recombinaison inhomogène le long du génome. Certains paquets de locus sont transmis intacts de générations en générations ; ces groupes de marqueurs sont appelés **blocs de LD** ou bloc haplotypiques ou encore haploblocs en raison du fort degré de LD et de la faible diversité haplotypique qui en résulte (Collins et al 1999, Nordborg and Tavaré 2002, figure 1.2-B p. 13). En conséquence, l'essentiel de l'information concernant le *pattern* de variation génétique au sein d'un bloc peut se résumer à partir d'un sous-ensemble de locus. Le projet *HapMap*[8] a pour objectif de décrire ces *patterns* de variation génétique commun chez l'Homme, en délimitant les blocs de LD le long du génome.

En dehors d'un taux de recombinaison inhomogène, on a montré qu'une structuration du génome en blocs du LD pouvait résulter d'autres phénomènes liés à la population, à son environnement et au génome lui-même : la dérive génétique[9], la croissance de la

[8] `http://www.hapmap.org`

[9] fixation allélique aléatoire au sein de la population

population, la stratification de population[10], la sélection naturelle, les mutations ainsi que les conversions de gènes (Zavattari et al 2000).

Équilibre d'Hardy-Weinberg

- Principe : l'équilibre d'Hardy-Weinberg est l'un des principes fondamentaux de la Génétique des Populations. Il prescrit que sous certaines conditions et après quelques générations, les proportions génotypiques d'un locus se fixent autour d'un équilibre : l'équilibre d'Hardy-Weinberg. Il spécifie aussi ces proportions génotypiques comme une simple fonction des proportions alléliques. Les conditions pour atteindre l'équilibre sont : **(i)** population infinie ou suffisamment grande pour minimiser les effets de la dérive génétique, **(ii)** population panmictique, c'est à dire que les accouplements se font de manière équiprobables, **(iii)** pas de sélection, pas de mutation, et pas de migration de population de façon à se prémunir des pertes ou gains d'allèles et enfin **(iv)** les générations successives sont discrètes. Dans le cas le plus simple d'un locus présentant deux allèles (a, A) avec les proportions p_a et $p_A = 1 - p_a$ respectivement, les proportions génotypiques (p_0, p_1 et p_2) à l'équilibre seront alors données par (figure 1.2-C p. 13) :

$$\begin{cases} p_0 &= p_a^2 \\ p_1 &= 2p_a p_A \\ p_2 &= p_A^2 \end{cases}$$

- Déviation par rapport à l'équilibre : lorsque les conditions d'Hardy-Weinberg ne sont pas respectées, les proportions observées peuvent dévier des valeurs attendues. Si les contraintes de panmixie et de population infinie affectent directement ces proportions, la migration, la sélection et la mutation changeront les proportions alléliques mais la population continuera à respecter les proportions génotypiques prédites à l'équilibre à chaque génération. Le coefficient de consanguinité ($\mathcal{F}$) peut-être vu comme une mesure de déviation par rapport à l'équilibre, rendant compte de l'excès ou du déficit d'hétérozygotes dans une population ; Wright (1921) a proposé un modèle permettant de spécifier les proportions génotypiques à partir des proportions alléliques et du coefficient de consanguinité quand la population ne suit pas l'équilibre. Dans le cas simple d'un locus à deux allèles (figure 1.2-D p. 13) :

$$\begin{cases} p_0 &= p_a^2 + \mathcal{F} p_a p_A \\ p_1 &= 2p_a p_A - 2\mathcal{F} p_a p_A \\ p_2 &= p_A^2 + \mathcal{F} p_a p_A \end{cases}$$

Un coefficient de consanguinité positif ($\mathcal{F} > 0$) induira un déficit d'hétérozygotes dans la population, un coefficient négatif ($\mathcal{F} < 0$) induira un excès. On peut également noter qu'un coefficient nul ($\mathcal{F} = 0$) correspond à une population à l'équilibre.

[10] la population d'intérêt comporte des sous-groupes d'individus qui sont en moyenne plus apparentés les uns aux autres qu'aux membres des autres sous-groupes

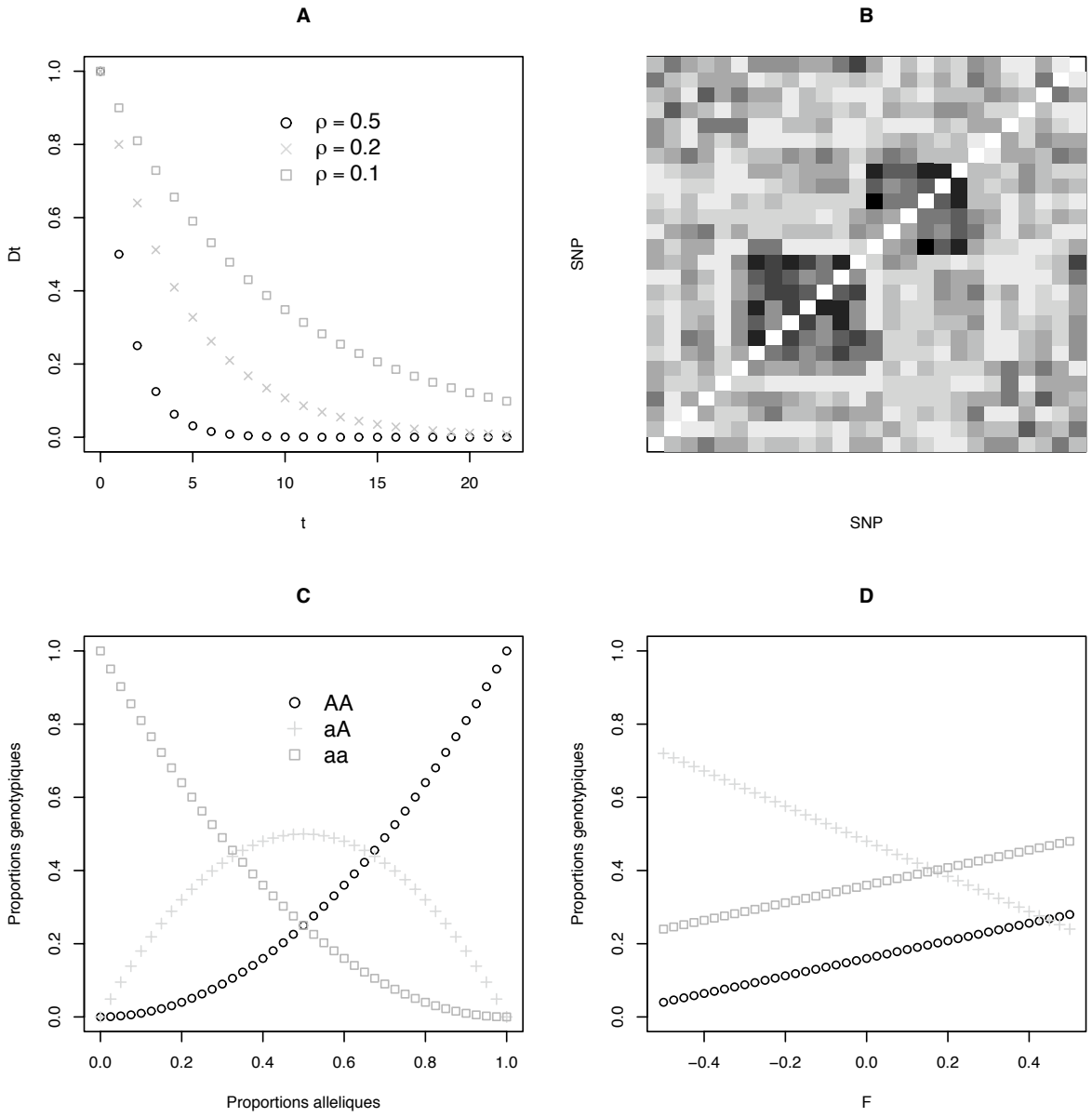


Fig. 1.2 – **A - Décroissance du LD :** le LD entre deux locus diminue avec le temps (t) et le taux de recombinaison (ρ), en fonction de la valuer initiale de LD au moment de la mutation ($\mathcal{D}_0$) : $\mathcal{D}_t = \mathcal{D}_0(1-\rho)^t$ (modèle de Malecot). **B - Blocs de LD** : on a représenté l'intensité du LD (r^2) entre chaque SNPs d'une région du chromosome 12 pour une population de Français, afin de mettre en avant les blocs de LD (en sombre). **C -** Proportions génotypiques en fonction des proportions alléliques à l'équilibre d'Hardy-Weinberg. **D -** Proportions génotypiques en fonction du coefficient de consanguinité ($p_a = 0.4$).

1.3 Épidémiologie Génétique

Définition

L'Épidémiologie est une discipline scientifique qui étudie ce qui est relatif à la santé d'une population : maladies, conséquences des maladies, causes de mortalité et éléments de "bonne santé". L'épidémiologie est avant tout utilisée pour agir sur la santé des populations, dans un souci d'évaluation des risques, de prévention et d'intervention ; elle ne peut généralement pas donner la preuve des mécanismes qui provoquent l'apparition des problèmes de santé ; en revanche, elle évalue la vraisemblance statistique d'une relation causale et permet d'agir sur la santé tout en se passant des contraintes expérimentales. Par exemple, on ne va pas inciter des sujets à fumer pour prouver une relation entre le fait de fumer et celui de développer un cancer du poumon. On distinguera : **(i)** l'Épidémiologie descriptive, qui décrit les phénomènes de santé en fonction des caractéristiques des individus tels que l'âge ou le sexe, **(ii)** l'Épidémiologie analytique, qui vise à établir les causes et l'ensemble des facteurs liés aux phénomènes décrits, **(iii)** l'Épidémiologie prospective qui vise à prévoir l'évolution des phénomènes pathologiques en s'appuyant sur des connaissances obtenues, et **(iv)** l'Épidémiologie d'intervention qui consiste à étudier les conséquences d'une méthodes de prévention sur le développement d'une maladie donnée.

On sait aujourd'hui qu'un grand nombre de pathologies a une composante génétique. L'Épidémiologie Génétique se définit donc comme la description et la compréhension de ces facteurs génétiques,

"a science that deals with etiology, distribution and control of disease in groups of relatives and with inherited causes of disease in population" (Morton 1998).

De fait, l'Épidémiologie Génétique se situe à l'interface de l'Épidémiologie et de la Génétique appliquée à l'échelle des populations. Des deux disciplines elle emprunte les objectifs, les problématiques ainsi que les outils d'analyse.

Étiologie simple, étiologie complexe

L'étiologie d'une maladie représente l'ensemble des mécanismes directement liés à son apparition. D'un point du vue génétique, dans le plus simple des cas, la modification d'un gène majeur est responsable à elle seule de l'apparition de la maladie. On parle alors de **maladie monogénique** ou d'étiologie simple. Il existe environ 6,000 maladies monogéniques répertoriées à ce jour. Parmi les plus courantes, on trouve par exemple la mucoviscidose, la drépanocytose, la chorée de Huntington, la myopathie de Duchenne ainsi que la dystrophie musculaire. Elles présentent en général un très fort caractère héréditaire.

Lorsque la maladie présente des composantes génétiques et environnementales multiples, on parle alors de **maladie multifactorielle** ou d'étiologie complexe. Leur mode de transmission est bien moins évident et la coexistence d'effets combinés de facteurs génétiques et environnementaux rend difficile la prédiction de la maladie au regard d'un gène seul. On fera alors plutôt référence aux gènes impliqués en tant que gènes de **prédisposition** ou de **susceptibilité**. Parmi les plus courantes, on pourra citer l'asthme, les maladies auto-immunes (*e.g.* diabète de type-I, sclérose en plaque, polyarthrite rhumatoïde), les maladies psychiatriques (*e.g.* schizophrénie) et les cancers (*e.g.* mélanome malin).

Des origines de la discipline à nos jours

- Les pères fondateurs : la paternité de la Génétique est souvent attribuée à Mendel (1822-1884). Il fut vraisemblablement le premier à utiliser le terme de "gène" dans le cadre de ses recherches sur la transmission des caractères héréditaires. Selon Mendel, les caractères se transmettent d'une génération à l'autre par le moyen de gènes qui suivent des lois précises de ségrégations, appelées **lois de Mendel**. Une telle idée était en totale opposition avec la théorie de l'époque qui misait plutôt sur un mélange équiprobable des caractères parentaux. Si les traits monogéniques suivent parfaitement les lois de Mendel et portent en conséquences - de façon un peu simpliste et abusive - aussi le nom de traits mendéliens, les traits complexes ont rapidement montré les limites de ces lois. Au début du 20ème siècle, des scientifiques tels que Galton (1822-1911) et Pearson (1857-1936) commencent à décrire des caractères qui ne semblent pas suivre les lois de Mendel. C'est en 1918 que Fisher (1890-1962) publie un traité sur les caractères polygéniques en introduisant le fait que le phénotype d'un individu peut résulter des effets conjoints de plusieurs gènes, aucun n'ayant par ailleurs d'effet majeur sur le caractère ; il marque ainsi le début de l'étude des traits complexes ou, par opposition aux traits mendéliens et de façon tout aussi abusive, traits non-mendéliens.

- L'impact de la Biologie Cellulaire et Moléculaire : l'Épidémiologie Génétique - et plus généralement la Génétique - a bien évidemment évolué en parallèle avec les progrès techniques en Biologie. Le Biologie Cellulaire a dans un premier temps permis de jeter les bases cellulaires de l'hérédité : le chromosome. L'avènement de la Biologie Moléculaire dans les années 1970 a permis ensuite l'élucidation des bases moléculaires de l'hérédité : la séquence d'ADN. Cela a accéléré le développement de marqueurs génétiques[11] à l'origine des premières tentatives de cartographie du génome vers la fin des années 1980. De fait, l'Épidémiologie Génétique englobe le concept d'Épidémiologie Moléculaire qui réfère à l'intégration des marqueurs génétiques dans les études épidémiologiques.

[11] séquence d'ADN variable dans une population dont la localisation est parfaitement connue

- L'Épidémiologie Génétique Moderne : appuyée par des projets internationaux de grande envergure (*Human Genome Project*, *HapMap Project*[12], dbSNP[13]) qui accompagnent l'accumulation d'une quantité importante de données, rendue possible par les récents développement technologiques en matière de génotypage, l'Épidémiologie Génétique réunit aujourd'hui tous les moyens nécessaires à l'élucidation des mécanismes génétiques des principales pathologies multifactorielles.

Les chances de réussites dans ce domaines devront s'appuyer parallèlement sur le développement de méthodes mathématiques, statistiques et informatiques permettant le traitement de l'ensemble des informations disponibles.

1.4 Cadres d'étude

Les marqueurs génétiques

Jusque dans les années 1980, les marqueurs utilisés en Épidémiologie Génétique étaient d'ordre biochimique. Il s'agissait dans un premier temps du groupe sanguin ABO (Race et Sanger 1975) puis de protéines porteuses de polymorphismes, c'est à dire présentent sous plusieurs formes identifiables grâce à leur différence de migration sur gel (Roychoudhuri et Nei 1988). Les avancées en Biologie Moléculaire ont ensuite permis le développement des marqueurs génétiques reposant sur la variabilité de la séquence d'ADN en des positions parfaitement connues[14]. Les premiers furent les *Restriction Fragment Length Polymorphisms* (RFLPs), les *Variable Number of Tandem Repeat* (VNTRs ou minisatellites) et les *Short Tandem Repeat Polymorphisms* (STRPs ou microsatellites).

Les plus récents sont les *Single Nucleotide Polymorphisms* (SNPs). Ils font partie de la famille des RFLPs, mais n'impliquent le changement (i.e. mutation, délétion, insertion) que d'un nucléotide en un locus donné et sont de manière générale bi-alléliques (a, A). Chaque individu sera donc porteur au niveau d'un SNP d'un des trois génotypes possibles, les deux génotypes homozygotes (aa et AA) et le génotype hétérozygote (aA ou Aa indiscernables l'un de l'autre). Les SNPs représentent une source d'information riche et abondante, apparaissant en moyenne jusqu'à une fois toutes les 2,000 bases le long des 3 milliards de lettres constituant le génome humain. La diminution à la fois du coût et du temps de séquençage a récemment ajouté à leur intérêt croissant. On peut aujourd'hui approcher à 0.0007 euros et 0.027 secondes le coût et le temps de génotypage d'un SNP par individu. Les SNPs localisés dans les régions codantes d'un gène peuvent par ailleurs jouer un rôle direct en altérant la forme et ainsi la fonction de la protéine produite à partir du gène en question (SNP *missense*).

[12] `http://www.hapmap.org`
[13] `http://www.ncbi.nlm.nih.gov/SNP`
[14] ou locus-spécifiques

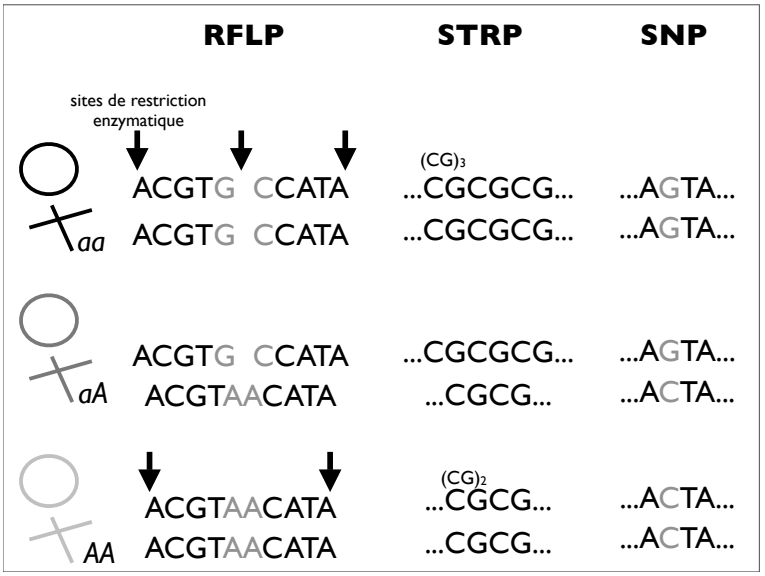


Fig. 1.3 – **Marqueurs génétiques :** on représente ici trois types de marqueurs génétiques (RFLP, STRP et SNP) bi-alléliques pour des individus homozygotes (aa et AA) et hétérozygotes (aA).

Données Familiales *vs* Cas-Témoins

- Études familiales : les études familiales traitent conjointement un certain nombre de familles de façon à détecter la transmission préférentielle d'allèles chez les personnes atteintes par la maladie considérée. Parmi les différents types d'études familiales, on distinguera celles reposant sur des ***pedigree*** (ou lignées familiales) plus ou moins larges de celles fondées sur des **familles nucléaires** (les deux parents et au moins un enfant) qui peuvent se réduire au cas des **trios** (les deux parents et un seul enfant). On peut de la même façon mener une étude sur la base de **fratries** (sans les parents), ce qui porte le nom d'étude de paires de germains ou *sib-pairs*.

- Études cas-témoins : aux études familiales on opposera une approche plus épidémiologique appelée cas-témoins. Une étude cas-témoins cherche à déceler une différence de distribution des variants génétiques entre une population de cas, constituée d'individus diagnostiqués avec la maladie d'intérêt, et une population de témoins sélectionnés dans la population générale et qui ne sont *a priori* pas porteurs de la maladie. A première vue, le recrutement de cas et de témoins peut paraître plus facile que celui de familles, du fait de la contrainte imposée dans ce dernier cas par l'obtention des génotypes d'individus apparentés. Il peut en effet s'avérer difficile d'obtenir les parents de chaque patient, en particulier lorsque la maladie se développe à un âge avancé. Cependant, le choix des cas et des témoins peut également soulever certains problèmes. En l'occurrence, il est nécessaire d'assurer l'homogénéité des deux groupes en prenant soin de les assortir sur des covariables

telles que l'âge ou le sexe, qui peuvent avoir une influence sur le phénotype observé et ainsi biaiser le facteur génétique que l'on cherche à détecter. Si certains biais comme le sexe ou l'âge sont aisément contrôlables lors du recrutement des individus, d'autres tels que l'origine éthnique le sont plus difficilement (voir Stratification p. 34).

Les études cas-témoins présentent par ailleurs une flexibilité qui leur permet de s'adapter aux facteurs de susceptibilités recherchés. Par exemple, plutôt que de s'intéresser aux déterminant génétiques impliqués dans le développement d'une maladie, on peut préférer rechercher ceux impliqués dans une forme de la maladie. Sur l'exemple du SIDA, l'idée est de rechercher les gènes responsables d'un développement lent ou rapide de la maladie chez les individus séropositifs. L'étude visera alors à contraster les phénotypes extrêmes (ici développement lent contre rapide), plutôt que d'utiliser des témoins séronégatifs (Hendel et al 1996).

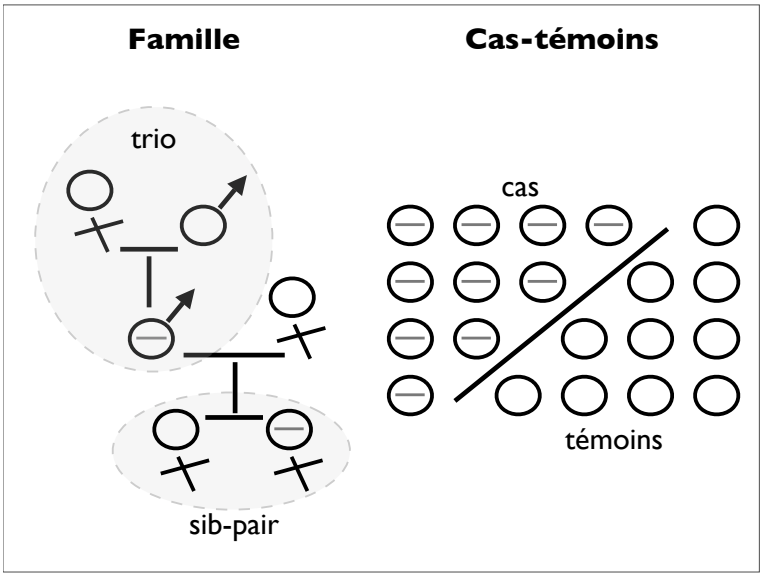


FIG. 1.4 – **Etudes familiales et cas-témoins.**

Liaison *vs* Association

Le choix entre une étude familiale ou cas-témoins dépend en partie de la quantité que l'on veut analyser : la liaison et/ou l'association.

- Études de Liaison : une étude de liaison vise à quantifier l'excès d'allèles identiques par descendance mendélienne que partagent des germains atteints. Un test de liaison (*e.g. Lod-Score Test*) comparera les proportions génotypiques observées à celles attendues sous l'hypothèse nulle que les recombinaisons entre locus sont équiprobables. Si les études de

liaison ont rencontré un certain succès dans le passé, la taille relativement limitée des familles ne permet pas de traiter de petites fractions de recombinaisons[15], en raison du manque d'individus recombinants dans les échantillons.

- Études d'association : les études d'association cherchent à déceler une association entre un variant génétique et la maladie, à un niveau populationnel et non plus familial uniquement. L'association peut-être directe si le marqueur observé est un locus de susceptibilité, ou indirecte si celui-ci se trouve physiquement proche du locus de susceptibilité et que leurs allèles sont statistiquement associés en raison du déséquilibre de liaison.

Les études d'association s'inscrivent naturellement dans un cadre cas-témoins : l'association entre variant génétique et statut cas-témoin peut facilement s'établir en utilisant des méthodes épidémiologiques classiques pour les études d'association cas-témoins (Breslow et Day 1982). Le chapitre 2 (p. 39) est dédié à cette thématique. En raison de la difficulté liée à l'obtention de témoins appropriés, le recrutement de parents représente une source idéale de témoins et plusieurs méthodes d'analyse d'association misant sur les familles ont été développées. Le *Transmission Disequilibrium Test* est l'exemple le plus connu : il teste l'association d'un allèle contre un autre dans un échantillon formé de cas et de leurs deux parents. Le génotype de chaque parent est considéré comme un couple de variables appariées - les allèles - dont l'un est transmis et l'autre non. L'hypothèse nulle d'indépendance est alors éprouvée en comparant la distribution des allèles transmis par les parents à celle des allèles non-transmis. Sous $H0$, on s'attend à ce que ces distributions coïncident. Le test de McNemar (1947) sur les variables appariées est adapté à ce problème et a permis d'aboutir au TDT (Spielman et al 1993). Un avantage prononcé du TDT et qu'il permet de détecter de la liaison seulement en présence d'association. En introduisant des données familiales, l'influence de potentielles associations liées à un manque d'homogénéité cas-témoins est éliminée. Néanmoins, lorsque la maladie se développe à un âge relativement avancé, il est quasiment impossible d'obtenir les génotypes parentaux. Une alternative raisonnable est alors d'utiliser de germains non affectés comme témoins. A partir du TDT, Spielman et Ewens (1998) ont proposé un test pour comparer les proportions alléliques entre des cas et leur germain non-affecté, appelé sib-TDT.

Lorsque l'on amorce une étude en Épidémiologie Génétique, l'on peut se demander quelle approche doit être choisie : liaison ou association ? Très tôt, Risch et Merikangas (1996) ont suggéré, à partir de simulations, que les études d'association sont vraisemblablement plus puissantes que les études de liaison pour identifier des effets modestes. Cette idée s'est rapidement installée dans l'esprit de la communauté ; la réponse n'est pourtant pas aussi tranchée. Il existe une différence d'échelle lorsque l'on considère la liaison ou l'association : **échelle de temps** tout d'abord puisqu'une étude de liaison s'intéresse à la transmission d'un marqueur avec le locus de susceptibilité sur quelques générations alors qu'une étude d'association repose sur l'association résultante d'un grand nombre de générations ; **échelle de distance** génétique ensuite puisque du fait de la faible fraction

[15] locus très proches

de recombinants observée dans une étude de liaison liée à la taille limitée des familles, celles-ci vont considérer la transmission de grandes régions génomiques avec le locus de susceptibilité (de l'ordre du cM[16]). Une étude d'association considère le résultat d'un grand nombre de recombinaisons et donc l'association entre un marqueur et le locus de susceptibilité s'opère sur de bien plus petites distances (de l'ordre de quelques dizaines à quelques milliers de paires de bases) ; **échelle de parenté** enfin puisque une étude de liaison focalise sur l'information contenue au niveau familiale alors qu'une étude d'association s'élève à la population entière. On peut d'ailleurs considérer la population comme une grande famille où les individus ont des liens de parenté plus faibles. On comprend donc qu'à l'origine, les études génétiques à grande échelle se faisaient sur la base de la liaison, nécessitant ainsi un nombre plus réduit de marqueurs pour couvrir l'ensemble du génome. Pour revenir à la différence de puissance entre les deux approches, Tu et Whittemore (1999) ont nuancé les propos tenus par Risch et Merikangas en suggérant que si, de façon générale, les études d'association étaient plus puissantes que les études de liaison, ce gain de puissance dépendait très fortement de paramètres tels que le déséquilibre de liaison entre les marqueurs utilisés et les locus de susceptibilité, ainsi que les proportions alléliques respectives. Dans certaines situations, ils ont observé que l'avantage de la liaison sur l'association n'était pas si net et que cette tendance pouvait quelques fois s'inverser.

On peut se poser la même question sur le choix de familles ou d'individus indépendants pour mener une étude d'association. Outre les considérations pratiques évoquées précédemment, les études cas-témoins souvent critiquées pour leur défaut de robustesse face à la stratification de population, semblent cependant généralement montrer plus de puissance que les études familiales, à taille d'échantillon équivalente (Morton 1998, Risch and Teng 1998).

Gènes-Candidats *vs Genome-Wide*

- Gènes-candidats : les approches gènes-candidats consistent à sélectionner un ensemble de gènes dont les fonctions pourraient intervenir dans l'étiologie de la maladie étudiée, et à les tester directement par association. Le choix des gènes peut être guidé par des *a priori* biologiques tels que la fonction ou l'appartenance à une voix métabolique associée à une maladie, ou encore sur la base de la localisation dans une région chromosomique d'intérêt, suggérée par une précédente étude de liaison ou d'association. Même lorsque les connaissances *a priori* sont larges et que la physiopathologie de la maladie est relativement bien comprise, l'approche gènes-candidats n'identifiera qu'une fraction des déterminants génétiques. Dans le cas contraire, elle est alors inadaptée pour appréhender de façon exhaustive les causes génétiques des maladies.

[16] le centimorgan noté cM est une unité de distance génétique. Elle représente la probabilité de recombinaison sur une distance : 1cM = une probabilité de 1% de recombiner

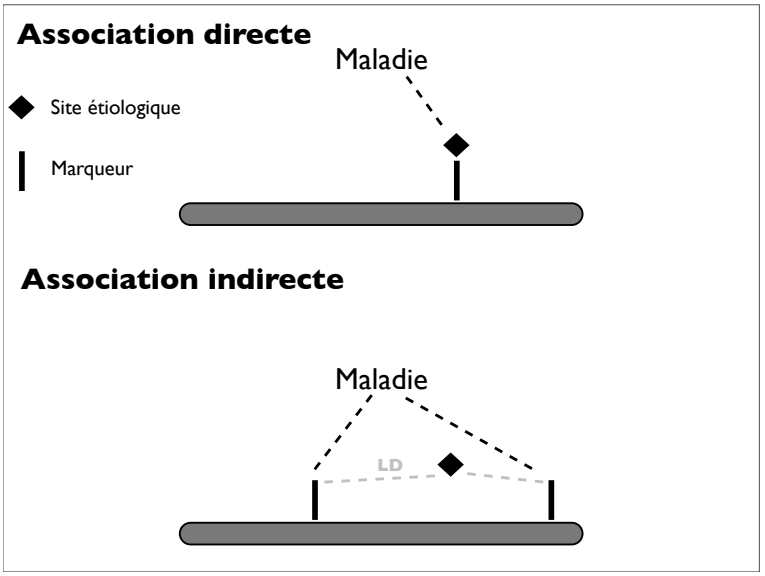


Fig. 1.5 – **Association directe et indirecte.**

- *Genome-Wide* : une étude d'association *genome-wide* (ou systématique) investit une grande partie du génome sans aucun *a priori* sur l'identité des locus impliqués. Cette approche représente une stratégie impartiale, non dirigée et assez complète pouvant être mise en place en l'absence d'indices sur la fonction ou la position des locus de susceptibilités. Elle a d'abord été utilisée pour des études de liaison utilisant des microsatellites et a permis de mettre en évidence la plupart des gènes responsables des maladies monogéniques connus. Malheureusement, cette approche a eu des difficultés à s'étendre aux maladies multifactorielles, l'excès de transmission chez des apparentés atteints étant plus faible pour des effets modérés. Les études d'association *genome-wide* sont donc apparues comme une alternative de choix et devaient constituer pour Risch et Merikangas (1996) l'avenir des études génétiques des maladies complexes, catalysées par l'ensemble des avancées technologiques.

Du choix de l'approche va dépendre la sélection des marqueurs. Mais il est évident que ce choix est contraint aux moyens dont dispose le laboratoire. De grands centres de séquençage et de génotypages[17] se forment afin de centraliser la technologie nécessaire au lancement d'études de grande envergure.

[17]par exemple le *Centre National de Génotypage*, Evry

1.5 Déroulement d'une étude d'association *genome-wide*

Étape 1 : Recrutement des individus et questions éthiques associées

- Recrutement : le recrutement des individus introduits dans l'étude constitue la première étape de toute étude épidémiologique et dépend du cadre choisi : familles ou cas-témoins. Le recrutement des cas nécessite un diagnostic précis de la pathlogie fondé sur des indicateurs histologiques, physiologiques, somatiques ou encore comportementaux dans le cas des maladies psychiatriques. Le diagnostic peut s'avérer plus ou moins évident en fonction de la nature de la pathologie et de la précision des indicateurs.

- Questions éthiques : la protection des individus est un aspect fondamental de toute étude épidémiologique. La mise en place de ressources génomiques publiques, à partir de populations et d'individus identifiés, soulève donc des questions scientifiques, sociales et éthiques qui sont inextricablement liées.

" I think you need to give conscious consent to having any data, any personal data used, whether you are identified or not. That's certainly a right. That's your information, it's your medical history. Whether it's identified or not, you should control it." Patient 14 dans Willison et al (2003).

A partir de l'exemple du projet *HapMap*[18] nous allons évoquer quelques uns des problèmes éthiques qui se posent aujourd'hui, et comment ce projet y a répondu de manière à concilier les différentes législations et sensibilités socio-culturelles des pays participants (*the International HapMap Consortium*, 2004).

Le projet *HapMap* a pour objectif de décrire les *patterns* de variation génétique communs chez l'homme, en identifiant les blocs de LD le long du génome ainsi que les tagSNPs correspondants, c'est à dire les SNPs permettant de résumer l'information contenue dans un bloc. Les chercheurs peuvent alors examiner l'ensemble du génome à partir d'un nombre restreint de tagSNPs (à peu près 500,000) au lieu d'étudier les 10 millions de SNPs qu'il contient, ce qui permet de réduire considérablement le coût des études *genome-wide*. Afin d'assurer la prise en compte des questions éthiques à chaque étape du projet, un groupe ELSI (pour *Ethical, Legal and Social Implications*) a été constitué.

La première question concerne bien évidemment la **protection des participants** : dans les données *HapMap*, on trouve, avec les génotypes de chaque individu, une indication sur le sexe ainsi que sur la population d'origine. En revanche il n'y a aucune donnée permettant de faire le lien avec les donneurs : si les centres qui récoltent les échantillons

[18] `http://www.hapmap.org`

conservent les identifiants, ceux-ci ne sont naturellement pas mis à disposition du public. Les données n'incluent par ailleurs aucune information médicale ou phénotypique sur les donneurs. Afin de renforcer la protection de l'identité, le projet a recruté plus de donneurs que nécessaire de telle manière que le donneur lui même ne sait pas s'il a participé à la version finale du projet. Il est donc très difficile de faire le lien entre les données génomiques disponibles sur *HapMap* et l'un des participants. Cela peut très hypothétiquement arriver si l'on obtient les génotypes d'une personne dont on pense qu'elle a participé au projet, et qu'on les compare avec les données disponibles. Donc si le risque pour les participants d'être identifiés n'est dans l'absolu pas nul, il est cependant réellement négligeable.

Le deuxième question est liée à l'**identification des populations** à partir desquelles proviennent les échantillons. Pour des intérêts scientifiques et éthiques évidents, il était nécessaire de choisir des populations diverses, qui trouvent leur origine en Europe, en Asie et en Afrique. Si le fait de préserver l'anonymat des participants était naturel, celui de l'origine des échantillons l'était moins : connaître la population d'origine de chaque échantillon permettrait dans les études futures de choisir le jeu de marqueurs le plus pertinent en fonction de la population considérée. Par ailleurs les populations seraient en réalité facilement identifiables. Donc plutôt que de laisser aux chercheurs la possibilité d'inférer, éventuellement à tort, l'identité des populations participantes au projet, il paraissait plus judicieux de rendre cette information disponible.

Enfin la question des **bénéfices** d'un tel projet ce pose également. Ils devraient directement contribuer à l'amélioration de la santé, bien que cela puisse prendre quelques années avant de se matérialiser. A plus court terme, les principaux bénéficiaires du projet ne seront pas les participants eux-mêmes mais les chercheurs et les industriels qui vont développer de nouvelles molécules, des tests diagnostiques et tout autre produit commercial à partir des recherches utilisant les données *HapMap*. En revanche, le projet lui même n'a pas vocation à générer des bénéfices.

Étape 2 : Sélection des marqueurs et techniques de génotypage haut-débit

Pour une approche gènes-candidats, le choix des marqueurs est principalement guidé par le choix des gènes inclus dans l'étude. Pour une approche *genome-wide*, plusieurs stratégies de sélection sont envisageables en fonction de différents *a priori* biologiques et techniques pris en compte.

- Sélection LD : pour être utile, un marqueur doit être lui-même un locus étiologique ou alors être en déséquilibre de liaison avec un locus étiologique (Kruglyak 1999, Jorde 2000). Comme nous l'avons déjà vu en abordant le déséquilibre de liaison (p. 9), le génome peut être réduit à un ensemble de blocs de LD dans lesquels chaque variant est fortement corrélé avec les autres (Daly et al 2001, Gabriel et al 2002). Un marqueur peut donc à

lui seul, porter l'information contenue dans la région en question : on parle de **tagSNP**. Sur la base du projet HapMap, on estime aujourd'hui que quelques centaines de milliers de SNPs bien choisis devraient suffire à résumer l'ensemble des variations génétiques du génome humain. Le nombre précis de tagSNPs nécessaires dépend en fait de la population et de la méthode employée pour les déterminer (Zhang et al 2002).

- Sélection *Missense* : vu le nombre important de mutations *missense* parmi les polymorphismes à la base des maladie monogéniques, Botstein et Risch (2003) ont proposé de se focaliser sur les SNPs *missense*[19] ; un gène contenant en moyenne un ou deux SNPs *missense* (Cargill et al 1999), cette stratégie implique le génotypage de 30,000 à 60,000 SNPs. Si les variants sérieusement associés à des maladies incluent effectivement une forte proportion de variants *missense*, cet argument est néanmoins biaisé par le fait que jusqu'à présent, ces derniers ont été préférentiellement analysés. Par ailleurs, il est assez vraisemblable que les allèles impliqués dans des maladies multifactorielles soient plutôt localisés au niveau de variants non-codants, impliqués dans des événements de régulation et avec un impact plus modeste sur l'expression des gènes. L'on peut donc s'interroger sur l'efficacité d'une telle stratégie de sélection.

- Sélection Gène-Centré : il s'agit ici de retenir les marqueurs sur la base de leur proximité avec un gène. Comme précédemment, la sélection peut-être accompagnée d'*a priori* biologiques tels que le déséquilibre de liaison ou la nature des variants (e.g. *missense*). Cette stratégie nécessite cependant la connaissance de tous les gènes présents dans le génome et fait totalement abstraction de l'implication d'éventuels éléments de régulation situés dans des régions dépourvues de gènes.

- Sélection Pragmatique : une dernière stratégie consiste à sélectionner les variants en fonction de considérations logistiques telles que la facilité ou le coût de génotypage. Récemment, de larges collections pouvant aller de 10,000 à 1,000,000 SNPs ont été développées et proposées à un coût tout à fait raisonnable. Ces collections permettent de couvrir une fraction significative du génome, bien qu'étant *a priori* moins exhaustives qu'une sélection reposant sur le LD.

- Techniques de génotypage haut-débit : les méthodes de génotypage sont multiples et se sont développées au fur et à mesure des avancées en Biologie Moléculaire et des progrès technologiques. En particulier, on trouve la PCR[20] à la base de la plupart des techniques de génotypage. Le développement de technologies haut-débit et peu coûteuses a permis de se tourner vers des études d'association à grande échelle. Il en existe plusieurs

[19] qui change la constitution de la protéine produite par le gène dans lequel le SNP se situe

[20] permet d'augmenter la quantité d'ADN à partir d'une très faible quantité de départ

types à ce jour, les principales étant fondées sur le principe de **puces-à-ADN** (ou *DNA-microarray*).

Le principe des puces-à-ADN met en relation les principes d'hybridation entres brins d'ADN par complémentarité des bases, de fluorescence en microscopie et de capture d'ADN sur des surfaces solides. Les principaux composants d'une puce sont : **(i)** le support sur lequel est fixé l'ADN cible, **(ii)** les ADN sondes et **(iii)** un système de détection qui enregistre et interprète le signal d'hybridation. Deux types de puces ont été élaborés par les industriels *Affymetrix* et *Illumina*. Utilisant tous deux le principe de fluorescence, ils divergent principalement sur la nature du support d'hybridation : pour la technologie *Affymetrix*, l'ADN cible est directement synthétisé sur des puces, alors que pour la technologie *Illumina*, l'hybridation se réalise sur des billes.

Les méthodes d'analyse de puce-à-ADN permettant de déterminer la configuration génotypique d'un SNP, reposent avant tout sur la probabilité pour le signal de fluorescence résultant, de correspondre à tel ou tel génotype. Si pour le généticien il est important de travailler sur des génotypes précis, il y a cependant, lors du passage des probabilités à la détermination des génotypes, une perte d'information évidente pour le statisticien. Cette perte d'information peut avoir pour conséquence la détermination erroné d'un génotype ou une indétermination sur sa valeur (valeur manquante). Ce problème est évoqué page 32. Si aujourd'hui encore, on a l'habitude de distinguer l'analyse du signal de fluoresence de celle de l'association des génotypes avec la maladie, l'on peut néanmoins insister sur le bénéfice qu'apporterait la réalisation de ces deux analyses conjointement, sans passer par une discrétisation en génotypes du signal de fluorescence afin de prendre en compte l'incertitude du génotypage.

Enfin, pour se donner une idée du coût et du temps de génotypage, le génotypage d'un individu avec une puce *Affymetrix* 500K revient à 350 euros et il faut compter environ deux semaines à un expérimentateur pour génotyper 96 individus.

Étape 3 : Analyse statistique et formulation d'hypothèses

- Pré-traitement : une étape préliminaire à l'analyse consiste à mettre en forme les données, à les ranger dans des fichiers ou dans des bases de données, et à les "nettoyer" de façon à minimiser les éventuelles erreurs. L'importance de ce pré-traitement ne doit pas être sous-estimée puisqu'il peut faciliter les analyses et contribuer à la qualité des résultats (voir Contrôle Qualité p. 28).

- Analyse simple-marqueur : une part importante des analyses statistiques consiste dans un premier temps à traiter les marqueurs un par un, afin d'identifier individuellement ceux qui sont - directement ou non - associés à la maladie. Un chapitre de ce manuscript est consacré à ce type d'analyse (chapitre 2 p. 39).

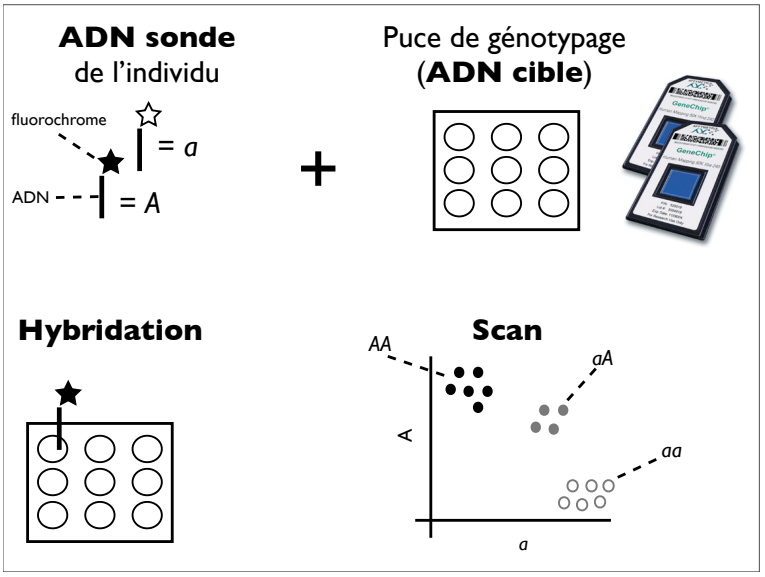


Fig. 1.6 – **Technique de génotypage haut-débit :** l'ADN de chaque individu est fragmenté et labelisé par un fluorochrome dont la couleur dépend de l'allèle présent sur la séquence d'ADN. Cette ADN sonde est ensuite déposé sur une plaque ou puce-à-ADN qui contient l'ADN cible (chaque puit représentant un locus donné) sur lequel il s'hybride. La puce est ensuite lue à l'aide d'un scanner et en fonction du signal de fluorescence observé on détermine le génotype de l'individu pour le locus considéré. Ici "noir" correspond au génotype homozygote AA, "blanc" au génotype homozygote aa, et un mélange des deux couleurs (gris) au génotype hétérozygote aA. Chaque locus est représenté plusieurs fois sur la puce de façon à améliorer la qualité des résultats.

- Analyse multi-marqueurs : en raison de la nature multifactorielle des maladies étudiées et du déséquilibre de liaison, il apparaît judicieux de prendre en compte les phénomènes d'interaction et d'association mis en jeu entre les marqueurs. Un autre chapitre développe cette thématique (chapitre 3 p. 103).

- Intégration de données hétérogènes : de la même façon qu'il est possible de choisir des gènes de susceptibilité *via* une approche gène-candidat, ou encore de sélectionner les marqueurs à génotyper en intégrant *a priori* des informations biologiques, il n'est pas exclu d'utiliser cette information *a posteriori* pour renforcer les hypothèses issues des analyses statistiques. On peut par exemple utiliser la nature génomique des marqueurs associés (*e.g. missense*), l'implication des gènes dans des voix métaboliques connues, l'homologie avec des gènes dont la modification dans des modèles animaux induit l'apparition de signes proches de la maladie étudiée, l'analyse d'expression différentielle de gènes au cours de processus physiopathologiques ainsi que l'information de similarité de séquences entre espèces dans le but de mettre en évidence des éléments de régulation.

Etape 4 : Vérification des hypothèses par réplication

En sciences expérimentales, la génération de connaissances nouvelles implique deux étapes distinctes : la production d'hypothèses et la vérification de ces hypothèses (Lantowski et Makalowski 2000).

En Épidémiologie Génétique, la réplication des résultats à travers une ou plusieurs populations indépendantes est considérée comme l'approche privilégiée pour cette vérification ; elle permet de distinguer les faux-positifs des vrais signaux d'association (Lander et Kruglyak 1995, Keightley et Knott 1999). Cependant la définition même de réplication n'est pas forcément claire et peut prendre un sens plus ou moins défini. De façon stricte, il s'agira de la réplication d'un même locus, mettant en cause les mêmes allèles ou génotypes de susceptibilité. On peut également considérer la *réplication technique* qui vise à réitérer la même expérience avec la même technique et sur la même population que l'on distingue de la confirmation technique qui réalise la même expérience sur la même population mais avec une technique différente, par exemple en utilisant une puce *Illumina* lorsque la première expérience a été faite à l'aide d'une puce *Affymetrix*.

En général, les résultats des réplications sont quelques peu décevants. Par exemple, sur plus de 1,300 études menées sur des maladies complexes et financées par le *National Insitutes of Health*, on estime entre 16% et 30% la proportion d'études mettant en avant des réplications et à quelques dizaines le nombre de variants génétiques identifiés comme étant impliqués dans une maladie complexe (Ioannidis 2003, Page et al 2003). On peut se demander pour quelles raisons les études d'association ne rencontrent pas le succès attendu ; les réponses peuvent se trouver en considérant la nature complexe des démarches expérimentales et des analyses statistiques mises en place, mais aussi des ma-

ladies auxquelles on s'intéresse. Dans ce contexte une attention toute particulière devrait être portée sur le contrôle qualité de l'ensemble d'une étude, de la génération des données à la formulation d'hypothèses.

1.6 Contrôle qualité : validité et fiabilité des résultats

L'un des moments les plus gratifiants pour le chercheur est certainement l'obtention des premiers résultats issus de son étude. Néanmoins, il doit immédiatement se poser la question *"Est-ce que je crois à ce que j'observe ?"*. La réponse à cette question est en grande partie déterminée par un certains nombre d'interrogations sous-jacentes concernant la qualité des données et des résultats obtenus.

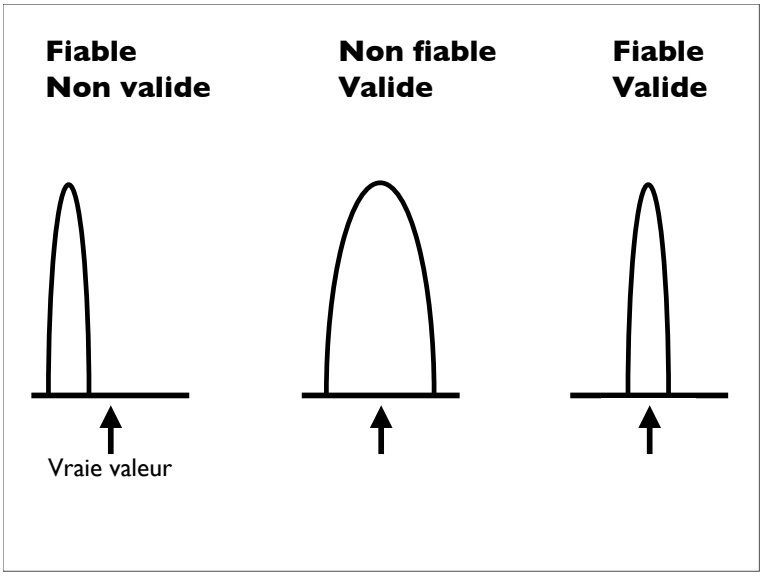


FIG. 1.7 – **Validité et fiabilité :** résultats d'un test d'hypothèse illustrant la distinction entre validité et fiabilité. La validité réfère à l'absence de biais dans un résultat, à sa justesse, tandis que la fiabilité représente sa reproductibilité ou précision.

L'obtention de données et de résultats totalement dépourvus d'erreurs est impossible. L'objectif n'est donc pas de n'avoir aucune erreur mais plutôt d'être capable de jauger l'étendue de tout type d'erreur, d'en estimer les conséquences et de les inclure dans l'interprétation des résultats.

Dans un article décrivant les causes de certains biais en Épidémiologie, Maclure et Scneeweiss (2001) suggèrent l'idée d'un "Episcope" à travers lequel un épidémiologiste observe l'association entre un agent causal et la maladie : de la même façon que l'utilisateur d'un télescope doit s'interroger sur l'existence et les conséquences d'une dégradation de l'image, un épidemiologiste devrait également s'interroger sur le pourquoi et le comment

ses résultats ont pu être dénaturés par des considérations liées à la qualité des données, l'efficacité des méthodes d'analyse ainsi que la validité et la fiabilité des résultats :

Les méthodes d'analyse employées sont-elles appropriées ? A partir de quel niveau de significativité doit-on conclure qu'un résultat est positif ? Les données ont-elles été correctement collectées ? La population utilisée était-elle appropriée ? Mon étude est-elle capable de saisir la nature complexe des données biologiques ?

Chacune de ces questions pointe un des principaux problèmes mis en cause dans la qualité des résultats. A l'occasion de cette section nous les aborderons un par un, nous en discuterons les effets et nous évoquerons les principales solutions existantes.

Puissance statistique

Le premier facteur à prendre en compte est le niveau de puissance statistique d'un test d'hypothèse ; il s'agit de sa capacité à rejeter l'hypothèse nulle quand celle-ci est effectivement fausse. En d'autres termes, la puissance d'un test est la probabilité de ne pas commettre une erreur de type-II :

$$\pi(\alpha) = \mathbb{P}_{H_1}(\mathcal{S} \geqslant t_\alpha) = 1 - \beta.$$

Le niveau de puissance atteint dépend de plusieurs paramètres de nature statistique ou spécifiques au problème génétique considéré. Ces paramètres sont :

- la statistique : la puissance d'un test dépend de la pertinence de la statistique elle-même et donc de son adéquation avec la question posée.

- le niveau α : la puissance est directement fonction du niveau fixé. Plus le taux d'erreur de type-I que l'on accepte de prendre est petit et plus la puissance sera faible. A l'inverse, une forte puissance sera accompagnée d'un taux d'erreur de type-I plus élevé, ce qui n'est évidemment pas souhaitable non plus. L'objectif est donc de trouver le compromis entre une puissance satisfaisante et taux d'erreur de type-I raisonnable.

- la taille de l'échantillon : plus l'échantillon est important et plus la puissance sera élevée. En Génétique, il y a fort à penser que dans la plupart des études publiées jusqu'à récemment, le manque de puissance avéré est essentiellement dû à des échantillons trop petits en comparaison des effets modestes que l'on espère identifier (Ioannidis et al 2001). A titre d'exemple, sur les 226 études d'association concernant l'ostéoporose, publiées au cours de l'année 2002, un peu plus de la moitié est fondée sur des échantillons de moins de 200 individus (Liu et al 2003). Des jeux de données de l'ordre du millier d'individus

apparaissent plus crédibles. Par ailleurs à taille d'échantillon constante, la puissance est maximale pour un nombre de cas et de témoins équivalents.

- la force d'association : plus l'association entre un variant génétique et la maladie sera forte, plus la différence entre ce que j'observe et ce que j'attends sous l'hypothèse $H0$ sera franche, et plus la puissance sera élevée.

- le déséquilibre de liaison : les études d'association reposent en partie sur le fait que le marqueur testé est en déséquilibre de liaison avec le variant étiologique. L'aptitude à identifier ces associations indirectes dépend du degré de déséquilibre de liaison qui existe entre les locus de susceptibilité et les marqueurs.

- les proportions alléliques : on sait que la puissance pour détecter une association dépend également des proportions alléliques du locus de susceptibilité et du marqueur dans le cas d'une association indirecte (Zondervan et Cardon 2004). En particulier, on a vu que le degré de déséquilibre de liaison (et donc la puissance) est maximum lorsque les proportions alléliques entre les deux locus sont égales (déséquilibre parfait).

L'effet conjoint de ces facteurs peut faire sensiblement varier la puissance statistique d'une étude. L'étude de puissance, c'est à dire la procédure qui vise à estimer la puissance d'une étude ou d'une méthode statistique donnée sous certaines conditions, est aujourd'hui une démarche tout à fait courante en Statistique appliquée à la Génétique. Elle permet par exemple d'estimer la taille d'échantillon nécessaire pour atteindre un certain niveau de puissance sous une hypothèse alternative définie. Elle permet également d'estimer la puissance d'une approche de façon à la comparer à d'autres approches concurrentes. Cette démarche statistique va être employée à plusieurs reprise au cours de ce manuscrit.

Test-multiple

Lorsque l'on réalise un certain nombre de tests (n) avec un critère de rejet des hypothèses nulles ($H0_1, ..., H0_n$) du type $p\text{v} \leqslant \alpha$, on obtient des résultats de quatre natures :

	H_0 non rejetée	H_0 rejetée	total
H_0 vraie	vn	fp	V
H_0 fausse	fn	vp	F
total	$n - R$	R	n

On note fp le nombre d'erreurs de type-I (ou faux-positifs) et fn le nombre d'erreurs de type-II (ou faux-négatifs). Si l'on réalise ces n tests de façon indépendante, on rejette classiquement $H0_i$ lorsque $\text{pv}_i \leqslant \alpha$. Mais lorsque l'on prend en compte l'ensemble de ces n tests, le nombre de faux-positifs obtenus par chance augmente avec n. Par exemple, si je réalise 100,000 tests à un niveau 5%, je m'attends à obtenir sous $H0$ en moyenne $fp = n\alpha = 5,000$ faux-positifs ; la proportion de faux-positifs devient alors conséquente comparée au nombre de vrais-positifs que l'on s'attend à trouver. Afin de minimiser cette inflation de faux-positifs, l'idée est de choisir un seuil de décision plus stringent qui ne sera plus fixé à partir du risque d'erreur de type-I attribué à chaque test, mais plutôt fondé sur le contrôle de quantités plus adaptées et qui reposent sur le nombre d'erreurs générées par l'ensemble des n tests.

- *Le Family-Wise Error-Rate* : le FWER est défini comme la probabilité de rejeter à tort au moins une hypothèse alors que toutes les hypothèses testées sont nulles :

$$\text{FWER} = \mathbb{P}_{H_0}(fp > 0).$$

On peut également le trouver sous le nom de *Global* ou *Genome-wide Significance Level*. Dans le cas de tests indépendants, on a :

$$\text{FWER} = 1 - \mathbb{P}_{H_0}(fp = 0) = 1 - (1-\alpha)^n \leqslant \max(n\alpha; 1).$$

Cette majoration proposée par Bonferroni (1892-1960) est très proche de la vraie valeur du FWER dans le cas de tests indépendants ou faiblement dépendants. Néanmoins, elle tend à s'en écarter lorsque les tests sont dépendants, ce qui peut-être le cas des études d'association en raison du déséquilibre de liaison. Une alternative empirique permet d'estimer le FWER par simulations de Monte-Carlo qui intègrent l'éventuelle dépendance entre les tests. Le principe consiste à générer un certains nombre de jeux de données sous $H0$ à partir des données observées, en permutant les phénotypes de façon à ce que ces jeux de données satisfassent l'hypothèse de non-association et conservent le *pattern* de LD observé dans le jeu de données initial. En réalisant un grand nombre de permutations, on peut ainsi approcher la valeur du FWER. Si cette démarche est conceptuellement simple, elle demande en revanche un temps d'exécution plus important.

- *Le False Discovery Rate* : bien qu'un contrôle du FWER à 5% soit très largement utilisé en science, il apparaît inapproprié pour les études génétiques à grande échelle car trop conservatif : si le nombre de faux-positifs en est largement diminué, le nombre de vraies découvertes l'est tout autant. La plupart des chercheurs n'est conceptuellement pas contre le fait d'accepter un taux de faux-positifs un peu plus élevé en échange d'une augmentation significative de la puissance. Comme alternative au FWER, on a donc proposé le FDR qui est la proportion attendue de faux-positifs parmi l'ensemble des positifs :

$$\text{FDR} = \mathbb{E}(Q),$$

avec $Q = \frac{fp}{R}$ si $R > 0$ et $Q = 0$ sinon. Le FDR dépend du niveau individuel α associé à chaque test, de l'hypothèse alternative ainsi que de la probabilité *a priori* pour chaque test d'être ou non sous $H0$. Les deux derniers paramètres n'étant pas directement accessibles, le contrôle du FDR peut paraître moins simple à mettre en place que celui du FWER. Comme les p-values sont uniformément distribuées entre 0 et 1 sous $H0$ et que la probabilité *a priori* d'être sous $H1$ est tellement petite qu'on peut la considérer comme nulle, Benjamini et Hochberg (1995) estiment par $\widehat{V} = n\alpha$ la proportion de faux-positifs et proposent d'utiliser la majoration :

$$\text{FDR} \leqslant \max\left(\frac{n\alpha}{R(\alpha)}; 1\right),$$

avec $R(\alpha)$ le nombre de positifs observés à un niveau α donné.

Le problème du test-multiple est repris dans le chapitre 2 page 87 à l'occasion duquel nous évoquons une troisième quantité permettant le contrôle du nombre de faux-positifs et ainsi de fixer le seuil de rejet de l'hypothèse nulle. Il s'agit du FDR Local.

Erreurs de mesure

La mesure des facteurs étudiés n'est pas toujours évidente à réaliser avec précision et des erreurs de mesure peuvent survenir pour différentes raisons comme la qualité des échantillons, la performance des machines utilisées pour effectuer les mesures ainsi que les pratiques de laboratoire. De fait l'on doit se préparer à travailler avec des données de plus ou moins bonne qualité. Dans le cas de variables discrètes, on parlera d'**erreurs de classification**. Les erreurs qui affectent chaque individu de la même manière sont appelées erreurs **non-différentielles**. Mais il peut arriver que le erreurs dépendent de la valeur d'une ou plusieurs covariables ; on parle alors d'erreurs **différentielles**. Cela arrive par exemple lorsque les individus de différentes familles sont traités de manières différentes ; les conséquences des erreurs dépendront alors de la famille à laquelle appartient chaque individu. Les erreurs non-différentielles sont connues pour provoquer une diminution de puissance. Les erreurs différentielles peuvent être à l'origine d'effets plus sévères en affectant la nature des relations entre les marqueurs étudiés et la maladie (Fleiss 1981, Ewen et al 2000).

- Erreurs de génotypage : les erreurs de classification qui affectent les génotypes sont appelées erreurs de génotypages. Dans les études cas-témoin, il a été constaté qu'une augmentation du taux d'erreur de génotypage de 1% nécessitait d'augmenter la taille de l'échantillon de 8% pour maintenir constants la puissance et le taux d'erreur de type-I (Gordon et al 2002). Il n'existe pas vraiment d'approche standard pour détecter les erreurs de génotypage et minimiser leur effets.

De **bonnes pratiques en laboratoire** peuvent y contribuer, par exemple en "ran-

domisant" les expériences : en mélangeant les cas et les témoins lors des expériences, on se prémunit ainsi d'erreurs différentielles liées au statut ou à des variables latentes telles que l'effet du support de la puce de génotypage ou encore de l'expérimentateur pouvant entraîner des erreurs différentielles.

Une autre idée qui vient à l'esprit est de **répéter les génotypages** un certain nombre de fois ; de cette façon il est possible de comparer les différentes répétitions afin d'identifier les individus pour lesquels un même marqueur présente des génotypes différents. Néanmoins il n'est pas exclu que certains problèmes liés à la nature même du marqueur n'engendrent une erreur à chaque répétition. Par ailleurs, si les incohérences entre les résultats de plusieurs répétitions permettent d'identifier les marqueurs problématiques, elles ne permettent en revanche pas vraiment d'affirmer quel génotype est le bon. Enfin, la multiplication des génotypages entraîne une augmentation non-négligeable du coût et du temps de génération des données. Une façon de se convaincre de la qualité globale des données peut-être de génotyper un sous-ensemble de marqueurs "contrôles" en utilisant une méthode de génotypage plus fiable que les méthodes classiques de génération de données haut-débit (puces-à-ADN), fondée sur le séquençage par exemple.

Une autre façon de mettre en avant des erreurs de génotypages est de **tester l'équilibre d'Hardy-Weinberg** chez les témoins. Une déviation par rapport à l'équilibre peut se produire par chance ou en raison d'événements génétiques et évolutionaires (p. 12) ; une déviation peut aussi être causée par des problèmes techniques liés au génotypage tels que la non-spécificité des sondes utilisées ou encore le typage systématique d'homozygotes en hétérozygotes et *vice versa*. Le test d'Hardy-Weinberg apparaît donc comme un moyen simple et efficace de discriminer les marqueurs soumis à un nombre important d'erreurs. Cette idée n'est en réalité pas nouvelle. Déjà dans les années 1970, des chercheurs testaient l'équilibre d'Hardy-Weinberg sur la base des groupes sanguins ; une déviation significative pouvait souligner l'évidence de complications expérimentales (Mourant et al 1976). L'efficacité du test d'Hardy-Weinberg dans ce contexte est cependant discutée. Des études de simulations montrent que ces erreurs ne génèrent pas assez de déviation pour être réellement détectées avec une puissance satisfaisante (Leal 2005, Cox et Kraft 2006) ; pourtant en pratique l'étude empirique proposée par Hosking et al (2004) tend à montrer l'inverse : sur 313 SNPs, 36 (soit 11.5%) dévient significativement de l'équilibre ; il a été *a posteriori* prouvé que 26 de ces déviations (soit 70%) trouvent leur origine dans des problèmes liés au génotypage. Par conséquent, en l'absence de méthodologie plus efficace, l'équilibre d'Hardy-Weinberg reste un moyen simple de mise en évidence d'erreurs.

- Erreurs de phénotypage : de la même manière qu'il existe des erreurs de génotypage, il peut exister des erreurs de classification liées à la détermination des phénotypes (Rice et al 2001, Egan et al 2003). Comme nous l'avons déjà évoqué le recrutement des cas nécessite un diagnostic précis de la pathologie fondé sur des indicateurs histologiques, physiologiques, somatiques ou encore comportementaux. Le diagnostic peut s'avérer plus ou moins évident en fonction de la nature de la maladie et de la précision des indicateurs. Comparées aux erreurs de génotypages, les erreurs de phénotypages ont reçu beaucoup

moins d'attention pour des conséquences pourtant similaires sur les résultats.

- Génotypes manquants : le problème lié au traitement des valeurs manquantes est courant en Statistique, en particulier lorsque les données proviennent d'individus ou d'expérimentations. Une première solution consiste à ignorer les valeurs manquantes en considérant qu'elles n'ont pas d'effet sur les résultats obtenus. Une telle démarche, bien qu'attractive, nécessite de poser un certains nombre d'hypothèses dont on ne peut être sûr qu'elles se réalisent en pratique. Une deuxième solution consiste à exclure les marqueurs pour lesquels les valeurs manquantes semblent être différentielles entre les cas et les témoins (Little et Rubin 1987). Cette idée part du principe que les marqueurs pour lesquels la proportion de valeurs manquantes ne diffère pas entre les cas et les témoins sont moins problématiques. Cela n'exclut pourtant pas une diminution de puissance et le fait que les valeurs manquantes se distribuent préférentiellement suivant une autre variable que le statut. On sait par exemple que dans l'utilisation de puces de génotypage, les génotypes hétérozygotes subissent une plus forte indétermination et donc un plus grand taux de valeurs manquantes que les génotypes homozygotes. Une troisième idée consiste à inférer les génotypes manquants plutôt que de les retirer de l'étude. Des algorithmes tels que l'EM (p. 193) et des procédures d'imputation de données (Rubin 1987) constituent un ensemble de réponses à ce problème. Les valeurs observées d'un même individu et/ou d'individus différents sont alors prises en compte pour imputer les valeurs manquantes ; on peut par exemple intégrer l'information apportée par le déséquilibre de liaison ou l'équilibre d'Hardy-Weinberg.

De façon générale, on peut penser que le taux d'erreurs de génotypages est étroitement lié à celui de génotypes manquants : les méthodes de génotypages pour lesquelles l'indétermination sur les génotypes est faible peuvent avoir tendance à commettre une erreur et réciproquement (Lamy et al 2006). Il y a donc dans les méthodes de génotypage haut-débit actuelles, un équilibre à trouver entre erreurs de mesures et valeurs manquantes.

Confusion et Stratification

"While the logical absurdity of attempting to measure an effect for a factor controlled by matching must be obvious, it is surprising how often investigators must be restrained from attempting this" (Mantel et Haenszel 1959).

On parle de **confusion** lorsque la prévalence[21] diffère d'un groupe d'exposition à un autre. Un facteur de confusion **(i)** permet de prédire la maladie en l'absence de la variable d'exposition (ici le génotype), et **(ii)** est associé à la variable d'exposition dans la population considérée. L'assortiment des cas et des témoins sur certaines covariables a pour objectif d'équilibrer le nombre de cas et de témoins dans des strates définies par

[21] risque de développer une maladie

ces covariables, de façon à éviter qu'elles ne jouent le rôle de facteur de confusion. On dit alors que l'on ajuste ou contrôle l'effet de ces covariables. Certaines covariables telles que le sexe ou la classe d'âge sont faciles à obtenir mais on peut s'attendre à des biais de sélection plus problématiques tels que l'origine ethnique des individus. La population est alors stratifiée, ce qui peut remettre en cause la validité d'une étude.

La stratification est donc la présence dans une population de plusieurs sous-groupes qui diffèrent en terme de prévalence face à la maladie ; toute maladie qui apparaît avec un risque plus élevé dans un de ces sous-groupes sera positivement associée à tout allèle dont la fréquence est plus importante dans ce sous-groupe. Elle peut résulter d'un mélange récent de populations ou d'un assortiment insuffisant des cas et des témoins. Un exemple amusant est donné par l'association entre les allèles de la région *HLA* et la capacité à manger avec des baguettes dans la population de San Francisco : les cas sont principalement d'origine asiatique alors que les témoins sont majoritairement d'origine caucasienne ; par ailleurs les proportions alléliques des polymorphismes de la région *HLA* sont différentes chez les caucasiens et les asiatiques ; par conséquent l'association observée n'a naturellement rien à voir avec le rôle du système *HLA* dans la dextérité manuelle nécessaire pour manger avec des baguettes.

La stratification peut être à l'origine de faux-positifs et d'une diminution de puissance (Deng 2001) mais son réel impact est sujet à discussion. Wacholer et al (2000) montre que ses effets ne sont pas aussi inquiétant qu'on l'avait imaginé, à moins que certaines conditions ne soient réalisées : les plus importantes sont une sensible variation en terme de proportion allélique et de prévalence entre les sous-groupes. En dehors de ces conditions les auteurs suggèrent que l'impact de la stratification a toutes les chances d'être minime, en particulier si le nombre de sous-groupes est supérieur à 2. Néanmoins, elle reste un facteur de confusion éventuel qu'il est important de pouvoir estimer et maîtriser.

Un moyen de se prémunir de la stratification est d'utiliser une **approche familiale** plutôt que cas-témoins. Mais aujourd'hui, un grand nombre d'études d'association est essentiellement fondée sur des cohortes d'individus non-apparentés. Des solutions ont donc été proposées, travaillant sur un jeu de marqueurs "neutres" face à la maladie.

- *Genomic Control* : il s'agit de l'approche la plus employée sans doute pour sa simplicité (Devlin et Roeder 1999). Elle consiste à estimer le degré de surdispersion (λ) de la statistique utilisée généré par la stratification, et de l'utiliser pour ajuster la statistique avant de réaliser le test. En pratique, il s'agit d'une normalisation empirique de la statistique par rapport à la distribution attendue sous l'hypothèse nulle. Cette approche est simple mais repose sur l'hypothèse que la stratification est constante le long du génome ce qui n'est pas forcément le cas, par exemple lorsqu'un locus subit une forte pression de sélection.

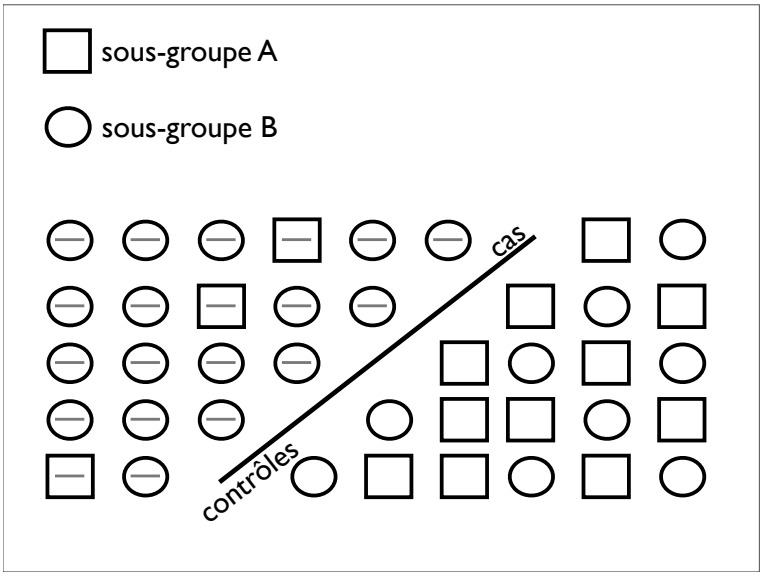


Fig. 1.8 – **Stratification :** la population est répartie en deux sous-groupes A et B. Ici les proportions des sous-groupes A et B sont différentes entre les cas et les témoins. Une différence de proportion allélique et de prévalence entre ces sous-groupes peut biaiser les résultats si la structure de la population n'est pas prise en compte dans les analyses.

- Approche régressive : une autre approche tout aussi simple et peut-être un peu plus naturelle d'un point de vue statistique consiste à introduire les marqueurs "neutres" en tant que covariables dans l'analyse, de façon à prendre en compte leurs effets et par la même occasion l'effet de la stratification (Setakis et al 2006).

- *Structured Association* : une troisième famille d'approches infère pour chaque individu le sous-groupe auquel il appartient et évalue l'association conditionnellement à cette structure estimée (Pritchard et al 2000, Satten et al 2001). Ces approches reposent sur des algorithmes d'inférence de variables latentes par des procédures de type EM (Dempster et al 1977) ou MCMC (Green et al 2003). Elles sont par conséquent un peu plus complexes à mettre en place, avec un temps d'exécution bien supérieur aux deux approches proposées précédemment.

- Autres approches : d'autres approches fondées par exemple sur des analyses en composantes principales (Price et al 2006) ou sur des modèles de mélange (Yu et al 2006) ont récemment été publiées et apparaissent comme des alternatives rapides et efficaces.

L'application de ces approches nécessite au préalable l'obtention d'un jeu de marqueurs "neutres". Le choix de ces marqueurs peut être problématique : si l'on se trouve dans le cas où les marqueurs à analyser sont sélectionnés sans *a priori* biologique, les marqueurs "neutre" par rapport à la maladie n'ont alors pas plus de raison de l'être que les marqueurs

utilisés pour l'analyse. De plus, en fonction de l'intensité de la stratification, le nombre de marqueurs neutres à inclure dans l'étude peut varier d'une dizaine à une centaine (Pritchard et al 2000).

Si la stratification est souvent mise en cause pour expliquer le manque de reproductibilité des résultats dans les études d'association cas-témoins, seulement un petit nombre d'études ont, jusqu'à présent, mis en évidence ses effets. Un exemple connu est l'association rapportée entre l'haplotype Gm3 ;5 ;13 ;1 avec le diabète de type-II (Knowler et al 1988) induite par un mélange dans l'étude de blancs européens avec des indiens américains Pima.

Étiologies hétérogènes

Au delà des problèmes que nous venons de voir et qui mettent en cause la qualité des données et des résultats, des conclusions contradictoires entre deux populations peuvent trouver leur origine dans de réelles différences biologiques. Lorsqu'une maladie implique des mécanismes différents en fonction des individus, on parle alors d'étiologie hétérogène. Cette propriété pose un certains nombre de défis pour la découverte de nouveaux locus de susceptibilité et leur réplication.

- Pléiotropie : la pléiotropie réfère à l'effet d'un même gène sur plusieurs phénotypes éventuellement pathologiques. Un exemple connu est celui donné par la phenylketonuria, maladie humaine responsable d'un retard mental et d'une dépigmentation de la peau mettant en cause le même gène (Campbell et Rudan 2002).

- Hétérogénéité allélique : on parle d'hétérogénéité allélique lorsque différents allèles confèrent la même pathologie. A titre d'exemple, on peut citer le cas du gène responsable de la fibrose cystique (Kerem et al 1989) pour lequel 75% des patients portent l'allèle Delta508 tandis que les 25% restant se partagent un grand nombre d'autres allèles.

- Hétérogénéité de locus : quand la maladie étudiée peut trouver son origine au niveau de différents locus, on parle d'hétérogénéité de locus. Un exemple classique est celui donnée par une forme récessive de l'albinisme qui peut être provoquée par un des deux gènes mis en cause (Trevor-Roper 1952).

- Conséquences et solutions : ces facteurs peuvent engendrer une population de cas inhomogène en termes d'étiologie et conduire à une diminution de puissance sensible. Les études familiales, en se concentrant sur une forme héréditaire d'une maladie et donc potentiellement plus homogène entre les cas, peuvent se montrer plus robustes face à ce

type d'hétérogénéité. Peltonem (2000) suggère de travailler avec des populations isolées[22] afin de minimiser le phénomène. Par ailleurs, d'un point de vue méthodologique, certains auteurs proposent de partitionner les cas en sous-groupes homogènes pour lesquels différentes étiologies sont autorisées (Whittemore et Halpern 2001, Province et al 2001, Ritchie et al 2003).

Hétérogénéité inter-populations : enfin ajoutons que tous ces facteurs peuvent agir différemment sur le même trait d'une population à l'autre. Bien que les variants étiologiques impliqués puisse être les mêmes, la variation du *pattern* de LD peut également accentuer les différences de résultats entre populations (Zavattari et al 2000).

La suite du manuscrit détaille plus spécifiquement le travail de recherche réalisé à l'occasion de cette thèse, inspiré des problématiques méthodologiques soulevées par l'analyse de données d'association cas-témoins *genome-wide*. Nous traitons successivement des approches simple-marqueur et multi-marqueurs.

[22] pour leur homogénéité

Chapitre 2

Approches simple-marqueur

La première étape de l'analyse statistique d'une étude d'association consiste souvent au traitement individuel de chaque marqueur. Ce chapitre est dédié à ce type d'analyse. Nous introduisons en premier lieu les notions d'association statistique, de test d'indépendance et de mesure d'association. Puis nous présentons les tests d'association génétique principalement utilisés dans la littérature : le test génotypique, le test de tendance, le test allélique et le test d'Hardy-Weinberg.

Au premier abord, les analyses simple-marqueur peuvent paraître tout à fait triviales ; elles soulèvent cependant un certain nombre de questions et de décisions à prendre en terme de stratégie d'analyse. En particulier, utiliser l'un ou l'autre des différents tests d'association peut changer la puissance de l'analyse ainsi que les résultats obtenus. Afin d'éclairer ce choix, nous mettons en place une étude de puissance. A cette occasion, nous comparons différentes méthodes d'estimation de la puissance. En l'occurrence, nous montrons que malgré une complexité avérée, l'utilisation de Formes Quadratiques présente un certain intérêt comparé aux méthodes traditionnelles. Ces résultats ont été publiés dans *Annals of Human Genetics* (2006). Nous avons également comparé la puissance des statistiques considérées. En particulier, nous consacrons une section au test allélique dont la validité dépend en réalité du respect de l'équilibre d'Hardy-Weinberg ; comme alternative, nous proposons un test allélique exact et valide en toutes circonstances que nous avons publié dans *Human Heredity* (2006). Nous consacrons également une section au test d'Hardy-Weinberg qui a récemment été proposé en tant que test d'association ; si sa validité dans le contexte des études d'association *genome-wide* est discutée, combiné à d'autre tests, il semble néanmoins apporter un gain de puissance tout à fait intéressant. Ce travail est actuellement en cours de soumission pour publication.

Enfin lorsque l'on réalise un grand nombre de tests, le problème du test-multiple est à considérer très sérieusement pour décider d'un niveau adéquat de rejet de l'hypothèse nulle. Cette thématique a été évoquée en introduction. Dans ce chapitre, nous abordons une quantité développée récemment, le FDR Local, pour lequel nous introduisons une méthode d'estimation aussi simple et rapide qu'intuitive.

2.1 Introduction

Les données issues d'études d'association cas-témoins ou familiales doivent indiquer pour chaque individu le statut (affecté/non-affecté) ainsi que la configuration génotypique de chaque marqueur. On peut également y trouver des informations complémentaires telles que les identifiants des individus, des marqueurs, la position des marqueurs sur le génome, les liens de parentés entre individus s'ils existent, ainsi que des covariables telles que le sexe, l'âge ou l'origine ethnique. Le format de données que nous utilisons pour intégrer toutes ces informations est réparti en trois tables : "geno", "pheno" et "info".

La table "geno" donne pour chaque marqueur (M_i), la configuration génotypique (0, 1 et 2 correspondant à aa, aA et AA) par individu (i_i) tandis que la table "pheno" indique pour chaque individu son phénotype (D pour affecté ou *diseased* et H pour non-affecté ou *healthy*) ainsi que les éventuelles valeurs de covariables (table 2.1).

	M_1	M_2	$\cdots$	M_n	statut	sexe	âge
i_1	0	0	$\cdots$	0	D	m	18
i_2	2	1	$\cdots$	1	D	f	39
i_3	2	1	$\cdots$	2	D	m	32
$\vdots$							
i_N	2	0	$\cdots$	1	H	f	35

Tab. 2.1 – **Geno et Pheno.**

La table "info" (2.2) donnent toutes les informations complémentaires sur les marqueurs, les deux principales étant le chromosome sur lequel ils se trouvent ainsi que leur position (en paires de bases par exemple).

	chr	position
M_1	1	11234
M_2	1	11889
M_3	2	436789
$\vdots$		
M_n	2	445631

Tab. 2.2 – **Info.**

Les approches simple-marqueur estiment l'effet marginal de chaque marqueur sur la maladie indépendamment des autres marqueurs. Il s'agit d'attribuer à chaque marqueur (M_i) une valeur de la statistique considérée ($\mathcal{S}_i$) et en fonction de cette valeur par rapport à un seuil déterminé par le taux d'erreur de type-I que l'on se donne (en général 5%), de décider si l'on considère ou non le marqueur comme étant statistiquement associé à la maladie (table 2.3).

	M_1	M_2	$\cdots$	M_i	$\cdots$	M_n
$\mathcal{S}$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\cdots$	$\mathcal{S}_i$	$\cdots$	$\mathcal{S}_n$
$H0$	non rejetée	rejetée	$\cdots$	non rejetée	$\cdots$	non rejetée

Tab. 2.3 – **Signal d'association.**

2.2 Association statistique et tests d'indépendance

Définition

Deux variables sont dites mutuellement dépendantes si la probabilité d'observer une valeur pour une variable dépend de la valeur prise par l'autre. L'association est une forme courante de dépendance et implique que le niveau général pris par une variable change en fonction des valeurs de l'autre. Il est important de noter que l'association n'implique pas forcément une relation de causalité : deux variables peuvent en effet être en association sans que l'une n'ait un effet direct et établi sur l'autre. Le concept d'association est sensiblement proche de celui de corrélation. Pour des variables quantitatives, la différence est que la corrélation implique une relation linéaire entre les variables alors que l'association n'est pas contrainte à la monotonie : une variable peut augmenter puis diminuer pendant que l'autre augmente. L'association est aussi le terme employé pour exprimer une dépendance entre variables qualitatives.

Une mesure d'association est une statistique qui permet de quantifier le degré de dépendance entre plusieurs variables. Un fort degré d'association indique que la connaissance du niveau d'une variable augmente fortement l'aptitude à prédire précisément le niveau de l'autre ; un degré peu élevé indique une moins bonne capacité de prédiction.

Table de contingence et modèles d'échantillonage

Soient deux variables qualitatives I et J pouvant prendre les valeurs $I_1, ..., I_p$ et $J_1, ..., J_q$ respectivement. La réalisation de N observations du couple de variables I, J peut se mettre sous la forme d'une table de contingence ($\mathcal{T}$), terme introduit par Pearson, où chaque case i, j représente le nombre d'occurrences du couple I_i, J_j (table 2.4).

L'ensemble des probabilités p_{ij}, $p_{i.}$ et $p_{.j}$ dénotent les distributions jointe et marginales de I et J respectivement. Les valeurs de chaque case sont issues d'un modèle d'échantillonnage donné en fonction de paramètres que l'on se fixe.

- Quand il n'y a aucune contrainte sur les valeurs prises par la table de contingence, un **modèle de Poisson** traite chaque case comme une variable aléatoire de Poisson de

	J_1	$\cdots$	J_j	$\cdots$	J_q	total
I_1	N_{11}	$\cdots$	N_{1j}	$\cdots$	N_{1q}	$N_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
I_i	N_{i1}	$\cdots$	N_{ij}	$\cdots$	N_{iq}	$N_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
I_p	N_{p1}	$\cdots$	N_{pj}	$\cdots$	N_{pq}	$N_{p.}$
total	$N_{.1}$	$\cdots$	$N_{.j}$	$\cdots$	$N_{.q}$	N

Tab. 2.4 – **Table de contingence ($\mathcal{T}$).**

paramètre $\lambda_{ij} = Np_{ij}$, indépendante des autres cases. La probabilité d'observer une table donnée est alors :

$$\mathbb{P}(\mathcal{T}) = \prod_i \prod_j \exp(-\lambda_{ij}) \frac{\lambda_{ij}^{Nij}}{N_{ij}!}.$$

- Quand la taille de l'échantillon est fixée mais que les marges en lignes et colonnes ne le sont pas, on applique un **modèle multinomial** $\mathcal{M}(N, p_{11}, \cdots, p_{ij}, \cdots, p_{pq})$ où chaque case correspond à un tirage possible sur les N réalisés. Dans ce cas, la probabilité d'une table satisfaisant $\sum_i \sum_j N_{ij} = N$ est donnée par :

$$\mathbb{P}(\mathcal{T}) = \frac{N!}{\prod_i \prod_j N_{ij}!} \prod_i \prod_j p_{ij}^{N_{ij}}.$$

- On peut également considérer que les marges en ligne ou en colonne sont fixées à l'avance. C'est le cas par exemple d'une étude cas-témoins où les nombres de cas et de témoins sont déterminés *a priori*. Dans ce cas chaque ligne est échantillonnée indépendamment des autres suivant un tirage multinomial $\mathcal{M}(N_{i.}, p_{i1}, ..., p_{ij}, ..., p_{iq})$. On parle alors de **modèle multinomial indépendant** ou de produit de multinomiales. Dans ce cas, la probabilité d'observer une ligne (l_i) satisfaisant $\sum_j N_{ij} = N_{i.}$ est :

$$\mathbb{P}(l_i) = \frac{N_{i.}!}{\prod_j N_{ij}!} \prod_j p_{ij}^{N_{ij}},$$

et la probabilité d'une table est donnée par :

$$\mathbb{P}(\mathcal{T}) = \prod_i \mathbb{P}(l_i) = \prod_i \left(\frac{N_{i.}!}{\prod_j N_{ij}!} \prod_j p_{ij}^{N_{ij}} \right).$$

- Enfin lorsque les marges en ligne et colonnes sont fixées, le modèle d'échantillonage approprié qui satisfait $\sum_j N_{ij} = N_{i.}$ et $\sum_i N_{ij} = N_{.j}$ est le modèle **hypergéométrique**

multivarié. La probabilité d'observer une ligne est alors :

$$\mathbb{P}(l_i) = \frac{\prod_{j=1}^{q-1} \binom{N_{ij}}{N_{.j}}}{\binom{N_{.j}}{N}},$$

et la probabilité d'une table est donnée par :

$$\mathbb{P}(\mathcal{T}) = \prod_{i=1}^{p-1} \mathbb{P}(l_i) = \prod_{i=1}^{p-1} \frac{\prod_{j=1}^{q-1} \binom{N_{ij}}{N_{.j}}}{\binom{N_{i.}}{N}}.$$

La distribution hypergéométrique telle qu'elle est souvent connue correspond au cas $i = j = 2$.

Tests d'indépendance

L'indépendance entre deux variables I et J correspond à l'égalité de chaque probabilité p_{ij} au produit des probabilités marginales : $p_{ij} = p_{i.} \times p_{.j}$. Ces probabilités n'étant en réalité pas connues, on utilisera leurs estimations au maximum de vraisemblance $\hat{p}_{i.} = \frac{N_{i.}}{N}$ et $\hat{p}_{.j} = \frac{N_{.j}}{N}$. Une statistique d'association quantifie alors la distance entre les valeurs observées de la table de contingence (N_{ij}) et celles attendues sous l'hypothèse nulle d'indépendance $(E_{ij} = N\hat{p}_{i.}\hat{p}_{j.} = \frac{N_{i.}N_{j.}}{N})$.

- Test de score : en 1900, Pearson introduit le test de score correspondant à ce test d'indépendance. La statistique associée quantifie l'écart normalisé entre les valeurs observées de la table de contingence et celles attendues sous $H0$:

$$X = \sum_i \sum_j \frac{(N_{ij} - E_{ij})^2}{E_{ij}}.$$

- Test du rapport de vraisemblance : une alternative au test de score proposé par Pearson est le test du rapport de vraisemblances :

$$\Lambda = \frac{\prod_i \prod_j (N_{i.}N_{.j})^{N_{ij}}}{N^N \prod_i \prod_j (N_{ij})^{N_{ij}}}.$$

En pratique on utilisera plutôt la statistique :

$$G = -2 \log \Lambda = 2 \sum_i \sum_j N_{ij} \log \frac{N_{ij}}{E_{ij}}.$$

Calcul de significativité : asymptotique, exact et empirique

Il existe plusieurs manières d'estimer la p-value associée à une valeur observée d'une statistique : **(i)** on connaît parfaitement la loi de probabilité que suit la statistique en question et l'on peut alors déterminer la valeur exacte de la p-value ; **(ii)** on ne connaît pas exactement cette loi mais on sait que sous certaines conditions, la statistique suit asymptotiquement une loi de probabilité connue ; **(iii)** lorsque l'on a n'a aucune indication exacte ou asymptotique sur la loi de probabilité suivie par la statistique, on peut l'approcher de façon empirique, en simulant un grand nombre de réalisations de la statistique en question sous $H0$. Il existe donc trois façons d'estimer la p-value d'une observation. Historiquement et encore aujourd'hui pour des raisons de simplicité, l'approche asymptotique reste la plus utilisée.

- Test asymptotique : sous $H0$, la statistique de Pearson (X) et la statistique du rapport de vraisemblance (G) suivent toutes deux asymptotiquement une distribution du χ^2 à k degrés de liberté. La p-value associée à la valeur observée de la statistique ($\mathcal{S}^{\text{obs}}$) est alors :

$$p\text{v} = \mathbb{P}(\chi^2(k) \geqslant \mathcal{S}^{\text{obs}}).$$

Le degré de liberté (k) correspond au nombre de paramètres libres à estimer. Dans le cas général, il faut estimer tous les p_{ij} en prenant en compte la restriction linéaire $\sum p_{ij} = 1$ ce qui donne $pq - 1$ paramètres à estimer. Sous $H0$, l'estimation des p_{ij} est déterminée par $p_{i.}$ et $p_{.j}$ avec les restrictions linéaires $\sum p_{i.} = 1$ et $\sum p_{.j} = 1$ soit $(1-p) + (1-q)$ paramètres à estimer. Il reste donc $(pq-1) - [(1-p)+(1-q)]$ soit $k = (p-1)(q-1)$ degrés de liberté.

En pratique, X et G tendent asymptotiquement vers la même distribution. A degré de liberté fixé, quand N augmente la distribution de X converge cependant plus vite vers un χ^2 que celle de G (Koehler et Larntz 1980). L'adéquation de l'approximation dépend du nombre de cellules (pq) et de la taille de l'échantillon (N). Cochran a étudié cette approximation dans plusieurs articles. En 1954 il propose que tout E_{ij} doit être supérieur à 5 mais que des valeurs plus faibles ne devraient pas poser de problème pour de grandes tables de contingences si au moins 4/5ème des E_{ij} sont supérieurs à 5. Il n'y a pas vraiment de règle générale mais supposer une distribution du χ^2 pour X ou G dans des cas où l'approximation n'est pas justifiée peut entraîner une imprécision dans l'estimation de la p-value, et par la même occasion une inflation ou une diminution du taux d'erreur de type-I.

- Test exact : un test exact consiste à un calcul exact de la p-value d'une observation. Dans le cas d'un test d'indépendance, il s'agit de calculer la probabilité d'apparition sous $H0$ d'une table de contingence ($\mathbb{P}_{H0}(\mathcal{T})$) en fonction du modèle d'échantillonnage considéré. La p-value de la valeur observée de la statistique ($\mathcal{S}^{\text{obs}}$) est alors déterminée

par la somme des probabilités des tables[1] pour lesquelles la valeur de la statistiques $(\mathcal{S}^{(i)})$ est au moins aussi extrême que $\mathcal{S}^{\text{obs}}$:

$$p\text{v} = \sum_{i|\mathcal{S}^{(i)} \geqslant \mathcal{S}^{\text{obs}}} \mathbb{P}_{H0}(\mathcal{T}^{(i)}).$$

Le nombre de tables possibles ainsi que leur probabilité dépendent du modèle d'échantillonnage choisi. Le plus courant est de considérer que toutes les marges sont fixées et que la table résulte d'un échantillonnage hypergéométrique (Yates 1934) ; le test exact est dit **conditionnel** dans le sens où la p-value est calculée conditionnellement à la valeur des marges que l'on se donne. Une autre façon de procéder consiste à considérer chaque ligne comme issue de tirages multinomiaux indépendants (Barnard 1945) ; dans ce cas seule une partie des marges sont fixées. Ce test est appelé test exact **non-conditionnel** par opposition au précédent, bien qu'il soit en réalité semi-conditionnel. Puisque les deux approches existent, la bonne manière de procéder a longtemps fait débat chez les statisticiens. Fisher critiqua longuement l'approche non-conditionnelle :

"(...) the existence of these less informative possibilities should not affect our judgment of significance based on the series actually observed. The fact that such an unhelpful outcome as these might occur (...) is surely no reason for enhancing our judgment of significance in cases where it has not occured; (...) it is only the sampling distribution of samples of the same type that can supply a rational test of significances".

Un test non-conditionnel permet de générer un plus grand nombre de tables possibles et réduit donc partiellement le problème de discrétisation qui est discuté par la suite (p. 46). Il a donc tendance à être moins conservatif que son homologue conditionnel. Il est en revanche beaucoup plus lourd en terme de temps d'exécution ce qui peut poser un vrai problème pour les tables de grande dimensions.

Le test exact le plus connu est le **test exact de Fisher** (1934). Il s'agit d'un test exact conditionnel. Le calcul de la p-value se fait en sommant les probabilités des tables dont la probabilité d'apparition est plus petite que celle de la table observée $(\mathcal{T}^{\text{obs}})$:

$$p\text{v} = \sum_{i|\mathbb{P}_{H0}(\mathcal{T}^{(i)}) \leqslant \mathbb{P}_{H0}(\mathcal{T}^{\text{obs}})} \mathbb{P}_{H0}(\mathcal{T}^{(i)}).$$

De fait, le test exact de Fisher n'est fondé sur aucune statistique claire et si les p-values qui en résultent tendent à être très proches de celles générées par un test exact conditionnel fondé sur la statistique de Pearson (X) ou du rapport de vraisemblance (G), l'on peut rencontrer en pratique quelques différences. Pour ces raisons, comme alternative à un test asymptotique fondé sur une statistique, nous aurons tendance à préférer le test exact, conditionnel ou non, mais correspondant à la même statistique. Par ailleurs le test exact requiert l'énumération d'un grand nombre de tables de contingence ce qui rend son temps d'exécution bien supérieur à un simple test asymptotique.

[1] parmi toutes les tables possibles $(\mathcal{T}^{(i)})$ en accord avec le modèle d'échantillonnage choisi

- Test empirique : lorsque la mise en place d'un test exact est fastidieuse ou impossible et que l'on ne connaît pas d'approximation asymptotique ou que celle-ci ne s'applique pas dans les conditions dans lesquelles nous nous trouvons, il est toujours possible d'approcher la distribution de la statistique par Monte-Carlo (Forster et al 1996). Cette approche empirique consiste dans notre cas à tirer un grand nombre de tables suivant le modèle d'échantillonnage considéré et à partir des paramètres d'échantillonnage déterminés par l'hypothèse nulle. A chaque simulation i, on obtient une valeur de la statistique $(\mathcal{S}^{(i)})$ et au bout de B simulations, une distribution empirique de la statistique $(\mathcal{S}^{(1)}, ..., \mathcal{S}^{(B)})$. La p-value est alors estimée par la proportion de valeurs simulées qui dépasse la valeur observée $\mathcal{S}^{\text{obs}}$:

$$\widehat{p\text{v}} = \frac{\sharp\{\mathcal{S}^{(i)} \geqslant \mathcal{S}^{\text{obs}}\}}{B},$$

où $\sharp$ est l'opérateur cardinal. Dans les études cas-témoins, une façon simple de procéder au test empirique conditionnel est de réaliser chaque simulation en permutant à chaque fois les statuts cas-témoins de façon à simuler sous l'hypothèse d'indépendance. Cette démarche est équivalente à un test empirique réalisé à partir d'un modèle d'échantillonnage hypergéométrique où toutes les marges sont fixées (figure 2.1).

Le calcul de la significativité par Monte-Carlo est facile à mettre en place mais requiert un temps d'exécution qui va dépendre de la précision que l'on souhaite atteindre pour l'estimation. Puisque $\widehat{p\text{v}} \times B$ est distribué suivant une binomiale $\mathcal{B}(B, p\text{v})$, $\widehat{p\text{v}}$ peut être approché par une distribution normale $\mathcal{N}\left(p\text{v}, \frac{p\text{v}(1-p\text{v})}{B}\right)$ qui donne pour $\widehat{p\text{v}}$ l'interval de confiance à 95% suivant :

$$\left[\widehat{p\text{v}} - 1.96\sqrt{\frac{\widehat{p\text{v}}(1-\widehat{p\text{v}})}{B}}; \widehat{p\text{v}} + 1.96\sqrt{\frac{\widehat{p\text{v}}(1-\widehat{p\text{v}})}{B}}\right].$$

Par conséquent l'ordre de grandeur de l'erreur de l'estimation de la p-value par Monte-Carlo décroit avec le nombre de simulations avec une vitesse $1/\sqrt{B}$. Le nombre de permutations nécessaires pour estimer avec précision des p-values très proches de 0 ou de 1 (de l'ordre de 10^{-8} par exemple) peut donc rapidement devenir conséquent et irréalisable en pratique lorsqu'il y a un grand nombre de tests à effectuer.

- Note sur la discrétisation : les statistiques d'association que nous considérons sont discrètes et de fait la p-value ne peut pas prendre toutes les valeurs comprises entre 0 et 1. En l'occurrence, il n'est généralement pas possible d'atteindre un niveau précisément égal à 5% pour le test et le vrai risque d'erreur de type-I est en pratique inférieur. En ce sens, l'approche traditionnelle du test d'hypothèse est conservative, c'est à dire que le test a tendance à ne pas suffisamment rejeter $H0$. Pour de grands échantillons, la distribution de la statistique est proche d'une distribution continue de type χ^2 donc la discrétisation ne pose pas réellement de problèmes. En revanche, pour les petits échantillons, sur lesquels on va appliquer un test empirique ou exact, la discrétisation de la statistique et donc des p-values est particulièrement marquée ce qui peut sensiblement diminuer le risque de première espèce et par la même occasion la puissance associée à un niveau fixé.

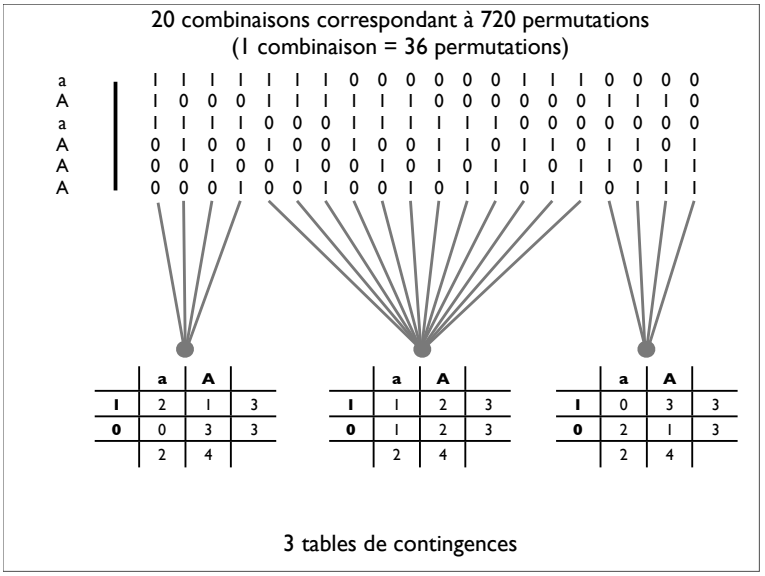


Fig. 2.1 – **Test empirique par permutation :** il revient à échantillonner suivant un modèle hypergéométrique où les marges lignes et colonnes sont fixées. La probabilité de tirer une permutation est de 1 sur le nombre de permutations possibles (ici 1/720, une combinaison correspondant à 36 permutations possibles). La probabilité de tirer une table correspond donc au nombre de permutations qui permettent de générer cette table divisé par le nombre de permutations possibles (ici 144/720, 432/720 et 144/720 respectivement) ce qui est strictement équivalent aux probabilités définies par le modèle hypergéométrique

Mesure d'association : *odds ratio*

Les statistiques proposées permettent de quantifier le degré d'évidence d'association et de réaliser un test mais ne sont en aucun cas des mesures d'association. En particulier elles dépendent du nombre d'observations (N). Pour différentes raisons l'*odds ratio* a émergé comme la mesure d'association la plus populaire en Épidémiologie. Soit p_x la probabilité d'être malade sachant que on est exposé à une certaine variable d'exposition (dans notre cas le génotype). L'*odd* (O) est alors défini comme le quotient de la probabilité pour un individu d'être malade sachant qu'il est exposé par la probabilité de ne pas être malade sachant qu'il est exposé :

$$O_x = \frac{p_x}{1-p_x}.$$

Les *odds* sont positifs ; ils sont supérieurs à 1 lorsque la chance pour un individu d'être atteint quand il est exposé est plus élevée que celle de ne pas l'être. Introduisons maintenant la probabilité pour un individu d'être malade sachant qu'il n'est pas exposé à la variable d'exposition ($p_{\bar{x}}$). L'*odds ratio* d'être atteint (OR) se définit alors comme le rapport des *odds* :

$$OR_{x,\bar{x}} = \frac{p_x(1-p_{\bar{x}})}{(1-p_x)p_{\bar{x}}}.$$

Historiquement, l'*odds ratio* était utilisé comme simple approximation dans les études cas-témoin du *risk ratio* ($\frac{p_x}{p_{\bar{x}}}$) longtemps considéré comme la mesure d'association de choix. Plus récemment l'*odds ratio* a gagné le statut de mesure d'association à part entière en raison de certaines propriétés intéressantes. Dans un premier temps, l'*odds ratio* peut être estimé sur un plus grand nombre de types d'études épidémiologiques que le *risk ratio*, c'est qui est un avantage pour comparer différentes études entre elles. Par ailleurs, il est symétrique de sorte que l'*odds ratio* d'être atteint est simplement l'inverse de l'*odds ratio* de ne pas être atteint :

$$\overline{OR} = \frac{1}{OR} = \frac{(1-p_x)p_{\bar{x}}}{p_x(1-p_{\bar{x}})}.$$

Enfin, et ce qui est peut-être la propriété la plus intéressante, l'*odds ratio* ne dépend pas de la taille de l'échantillon contrairement aux statistiques de Pearson et du rapport de vraisemblance introduites précédemment.

En toute généralité, la table de contingence du couple de variables I,J et de dimension $p \times q$ met en jeu un grand nombre d'*odds ratios* ; en réalité un par couple i,i' de valeurs prises par I croisé avec un couple de valeurs j, j' prises par J, soit $\binom{p}{2}\binom{q}{2} = \frac{pq(p-1)(q-1)}{4}$ *odds ratios* différents :

$$OR_{(ii'jj')} = \frac{p_{ij}p_{i'j'}}{p_{i'j}p_{i'j}}.$$

L'hypothèse d'indépendance dans une table de contingence revient à ce que tous les *odds ratios* valent 1. Ceux-ci s'estiment par régression logistique qui constitue aujourd'hui une approche de régression sur des variables binaires (cas-témoins par exemple) très populaire. Le lecteur intéressé par une énumération des autres mesures d'association peut se référer à Agresti (2002) pour plus de détails.

Régression Logistique

- Définition : la régression logistique est sans doute le modèle le plus important d'analyse de variables qualitatives binaires. Elle est largement utilisée dans un grand nombre de secteurs, et tout particulièrement en Épidémiologie où elle permet d'estimer les *odds ratios*. Ici la variable d'exposition est x ; soit p_x la probabilité d'être malade sachant la valeur de la variable explicative x. Il peut être tentant à première vue d'ajuster linéairement les valeurs prises par p_x sur celles de x. Néanmoins il est clair qu'une telle mise en relation peut mener à des valeurs de p_x en dehors de l'intervalle [0,1]. Il est donc nécessaire d'appliquer une transformation à p_x qui permet de contraindre ses valeurs entre 0 et 1 ; cette transformation est naturellement réalisée par la fonction logistique :

$$p_x = \text{expit}(\alpha + \beta x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

Ce modèle est appelé modèle de régression logistique en raison de la fonction logistique (ou expit) sur laquelle il repose. Quelles que soient les valeurs prises par α, β et x, p_x prend des valeurs comprises entre 0 et 1. Pour mieux comprendre les implications du modèle logistique, il peut-être intéressant de l'exprimer en fonction des *odds* :

$$O_x = \frac{p_x}{1 - p_x} = \exp(\alpha + \beta x).$$

Cette formulation montre que le modèle logistique est équivalent à un modèle exponentiel des *odds*. En considérant que la fonction logit d'un risque est définie par le logarithme des *odds* et en appliquant le logarithme de chaque côté de la formulation précédente, on obtient :

$$\text{logit}(p_x) = \log(O_x) = \log\left(\frac{p_x}{1 - p_x}\right) = \alpha + \beta x.$$

Sous cette forme équivalente, le modèle logistique trouve aussi le nom de modèle log-linaire des *odds* ou plus simplement modèle logit.

- Cas particulier des modèles linéaires généralisés : Le modèle logistique est en réalité un cas particulier des modèles linaires généralisés (GLM) dont la forme générale est :

$$\mathbb{E}(y|x) = f(\alpha + \beta x),$$

où f est une fonction continue et strictement croissante et g la fonction inverse de f telle que :

$$g(\mathbb{E}(y|x)) = \alpha + \beta x.$$

g est appelée fonction de lien et $\alpha + \beta x$ est le prédicteur linéaire. Pour le cas particulier du modèle logistique, f est la fonction expit et la fonction g de lien est la fonction logit. Pour plus d'information sur les GLM voir McGullagh and Nedler (1989).

- Interprétation des paramètres α et β : comment pouvons-nous interpréter β ? Son signe détermine si p_x augmente ou décroît avec x. Plus β est proche de 0 et plus p_x est insensible aux variations de x. Pour $\beta = 0$, p_x et x sont indépendants. Par ailleurs, pour chaque accroissement de x d'un unité, β a un effet additif sur l'échelle du logit et multiplicatif sur l'échelle des *odds*. En l'occurrence, pour deux valeurs x_A et $(x_A + 1)$ de x :

$$\begin{aligned} OR_{x_A+1,x_A} &= e^{\beta} \\ \text{logit}(p_{x_A+1}) - \text{logit}(p_{x_A}) &= \beta, \end{aligned}$$

et pour deux valeurs x_A et x_B de x :

$$\begin{aligned} OR_{x_A,x_B} = \frac{e^{\alpha+\beta x_A}}{e^{\alpha+\beta x_B}} &= e^{\beta(x_A - x_B)} \\ \text{logit}(p_{x_A}) - \text{logit}(p_{x_B}) &= \beta(x_A - x_B). \end{aligned}$$

L'interprétation du paramètre α a généralement moins d'intérêt ; il s'agit d'un paramètre de centrage du prédicteur linéaire sur 0 et peut s'interpréter comme la valeur du logit pour la valeur moyenne de x.

- Estimation : sous l'hypothèse que les observations sont indépendantes et pour un échantillon assez large, les valeurs des paramètres au maximum de vraisemblance ($\hat{\alpha}$, $\widehat{\beta}$) sont des estimateurs non-biaisés des vraies valeurs des paramètres (α,β). La fonction log-vraisemblance pour les modèles de régression logistique est strictement concave ce qui implique l'existence et l'unicité de $\widehat{\alpha}$ et $\widehat{\beta}$, excepté dans certains cas limites (Wedderburn 1976).

- Test de Wald : tester l'indépendance entre p_x et x revient à tester que l'*odds ratio* $e^{\beta} = 1$, ce qui revient ici à tester $H0 : \{\beta = 0\}$. D'après les résultats de Wald (1943) sur la distribution des estimateurs au maximum de vraisemblance, $\widehat{\alpha}$ et $\widehat{\beta}$ ont une distribution normale. Tester $H0$ peut donc revenir à réaliser un test de Wald sur β :

$$\mathcal{Z} = \frac{\widehat{\beta}}{\sqrt{\mathbb{V}(\widehat{\beta})}} \underset{H0}{\sim} \mathcal{N}(0,1).$$

Pour plus de détails sur les modèles logistiques voir Agresti (2002) ; pour plus de détails sur l'application de ce type de modèles à l'épidémiologie, se référer à Ahrens and Pigeot (2004).

2.3 Tests d'association marqueur-maladie

Table de contingence marqueur-maladie

Pour chaque marqueur, on établit une table de contingence à partir des données. Si cette démarche paraît simple, elle pose néanmoins une question sur la manière de répertorier les cas et les témoins. De façon assez naturelle, cela peut se faire sur la base des génotypes, ce qui conduit à une table de contingence génotypique $\mathcal{T}_G$ (2.5).

	aa	aA	AA	total
cas	D_0	D_1	D_2	N_D
témoins	H_0	H_1	H_2	N_H
total	N_0	N_1	N_2	N

Tab. 2.5 – **Table génotypique ($\mathcal{T}_G$).**

Une autre façon de procéder est de classer les individus en fonction de leurs allèles ; dans ce cas, chaque individu contribue à deux observations ce qui conduit à une table de contingence allélique $\mathcal{T}_A$ (2.6).

	a	A	total
cas	$D_a(= 2D_0 + D_1)$	$D_A(= 2D_2 + D_1)$	$2N_D$
témoins	$H_a(= 2H_0 + H_1)$	$H_A(= 2H_2 + H_1)$	$2N_H$
total	$N_a(= 2N_0 + N_1)$	$N_A(= 2N_2 + N_1)$	$2N$

Tab. 2.6 – **Table allélique ($\mathcal{T}_A$).**

Un comptage reposant sur les allèles peut paraître plus intéressant puisque la taille d'échantillon en est doublée. Mais les allèles n'agissent pas forcément de façon indépendante ce qui peut rendre un telle approche contre-intuitive et inappropriée.

A partir de ces tables de contingence, un certain nombre de tests d'association a été proposé, fondés sur les génotypes ou les allèles.

Tests fondés sur les génotypes

- Test génotypique : le test génotypique compare les valeurs observées dans la table de contingence génotypique à celles attendues sous $H0$:

$$H0 : \left\{ \begin{array}{lcllcl} p_{D_0} & = & p_D p_0 & ; & p_{H_0} & = & p_H p_0 \\ p_{D_1} & = & p_D p_1 & ; & p_{H_1} & = & p_H p_1 \\ p_{D_2} & = & p_D p_2 & ; & p_{H_2} & = & p_H p_2 \end{array} \right\}$$

Le test de score correspondant est fondé sur la statistique de Pearson :

$$X_G = \sum_{i=0}^{2} \frac{\left(D_i - \frac{N_D N_i}{N}\right)}{\frac{N_D N_i}{N}} + \frac{\left(H_i - \frac{N_H N_i}{N}\right)}{\frac{N_H N_i}{N}} \underset{H0}{\sim} \chi^2(2).$$

Ce test est le plus général. Il correspond au modèle de régression logistique où la variable explicative x est une variable qualitative pouvant prendre trois valeurs correspondant aux trois génotypes possibles. Ces trois possibilités sont codées dans le modèle sous la forme de deux variables[2] (x_1 et x_2) prenant les valeurs 0 ou 1 en fonction du génotype observé. En l'occurrence, le génotype aa sera codé par $\binom{x_1=0}{x_2=0}$, aA par $\binom{x_1=1}{x_2=0}$ et AA par $\binom{x_1=1}{x_2=1}$.

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2,$$
$$H0 : \{\beta_1 = \beta_2 = 0\}.$$

- Test de tendance : Cochran (1954) et Armitage (1955) ont été parmi les premiers à mettre l'accent sur l'avantage de prendre en compte l'ordre des variables lorsque celui-ci est justifié. Il suppose une relation linéaire entre les probabilités d'être malade sachant le génotype (p_{D_0}, p_{D_1} et p_{D_2}) ce qui restraint l'hypothèse alternative et permet de gagner 1 degré de liberté :

$$H0 : \{p_{D_0} = p_{D_1} = p_{D_2}\}$$

Le test de score correspondant est fondé sur la statistique :

$$X_T = \frac{N\left[N(D_1 + 2D_2) - N_D(N_1 + 2N_2)\right]^2}{N_D N_H\left[N(N_1 + 4N_2) - (N_1 + 2N_2)^2\right]} \underset{H0}{\sim} \chi^2(1).$$

Ce test correspond à un modèle de régression logistique où les génotypes sont codés par une variable qualitative x prenant les valeurs 0, 1 et 2 en fonction du génotype observé :

$$\text{logit}(p) = \alpha + \beta x$$
$$H0 : \{\beta = 0\}.$$

Ce modèle suppose un effet additif des allèles. Pour Slager and Schaid (2001) un des avantages du test de tendance est sa flexibilité puisqu'en faisant varier les valeurs prises par x, il est possible de traiter une grande variété de modèles différents (dominant, récessif... voir p. 57). Par ailleurs le test d'Armitage est le test utilisé par Devlin et Roeder (1999) pour contrôler les effets de la stratification à travers le *Genomic Control*.

[2] appelées facteurs ou *dummy variables*

Test d'Hardy-Weinberg : Il a été suggéré que de tester l'équilibre d'Hardy-Weinberg dans une population de cas pouvait aider à déceler de l'association (Feder et al 1996).

$$H0 : \left\{ \begin{array}{lcl} p_{D_0} & = & p_{D_a}^2 \\ p_{D_1} & = & 2p_{D_a}p_{D_A} \\ p_{D_2} & = & p_{D_A}^2 \end{array} \right\}$$

Le test de score correspondant utilise la statistique :

$$X_{\mathrm{HW}} = \frac{[D_0 - \frac{(2D_0+D_1)^2}{4N_D}]^2}{\frac{2D_0+D_1}{4N_D}} + \frac{[D_1 - \frac{(2D_0+D_1)(2D_2+D_1)}{2N_D}]^2}{\frac{(2D_0+D_1)(2D_2+D_1)}{2N_D}} + \frac{[D_2 - \frac{(2D_2+D_1)^2}{4N_D}]^2}{\frac{2D_2+D_1}{4N_D}} \underset{H0}{\sim} \chi^2(1).$$

De façon tout à fait équivalente, on peut fonder le test sur le coefficient de consanguinité observé chez les cas avec la statistique $F = \sqrt{N_D}\mathcal{F}$ avec : $F \underset{H0}{\sim} \mathcal{N}(1)$, en remarquant que $F^2 = X_{\mathrm{HW}}$.

Ce test n'est d'un point de vue statistique pas un vrai test d'association puisqu'il repose sur le déséquilibre d'Hardy-Weinberg engendré par une association entre le marqueur et la maladie. Du fait de cette particularité, le test possède des propriétés qui lui sont propres et qui sont discutées dans une section de ce chapitre (p. 77).

Tests fondés sur les allèles

- Test allélique : De façon similaire au test génotypique, il est possible de réaliser un test d'indépendance sur la table de contingence allélique :

$$H0 : \left\{ \begin{array}{lcl;{\quad}lcl} p_{D_a} & = & p_D p_a & p_{H_a} & = & p_H p_a \\ p_{D_A} & = & p_D p_A & p_{H_A} & = & p_H p_A \end{array} \right\}$$

La statistique de Pearson correspondante est :

$$X_A = \frac{2N((2D_2+D_1)(2H_0+H_1) - (2D_0+D_1)(2H_2+H_1))^2}{2N_D N_H (2N_2+N_1)(2N_0+N_1)} \underset{H0}{\sim} \chi^2(1).$$

Ce test correspond à un modèle de régression logistique où chaque individu apparaît deux fois et la variable explicative x prend les valeurs 0 si l'allèle observé est a et 1 s'il s'agit de A :

$$\mathrm{logit}(p) = \alpha + \beta x,$$

$$H0 : \{\beta = 0\}.$$

La justification de ce test dans les études cas-témoins est discutée puisqu'il traite les observations de façon indépendante. Or un individu contribue à deux observations. Cette omission statistique peut introduire un biais dans l'estimation de la p-value lorsque l'on réalise le test d'indépendance. Nous développons cette problématique et proposons une solution dans une prochaine section de ce chapitre (p. 69).

Méta-statistiques et combinaisons de tests

Des tests d'association combinant différentes statitiques ou différents tests on récemment fait leur apparition dans la littérature (Hoh et al 2001, Song et al 2005). Il peut s'agir de simples sommes, produits ou combinaisons linéaires de statistiques dans le cas des méta-statistiques ou bien encore de combinaisons booléennes de tests du type *on rejette* $H0$ *si tous les tests rejettent* $H0$ ou *si au moins un des tests rejette* $H0$.

La distribution de la plupart de ces combinaisons est facile à déterminer dans le cas de statistiques ou de tests indépendants. Par exemple la somme de deux χ^2 indépendants à k et l degrés de liberté est un χ^2 à $k+l$ degré de liberté. De la même façon, la probabilité de rejeter $H0$ pour deux tests indépendant est simplement le produit des probabilités de rejeter $H0$ pour chaque test. Le problème vient du fait que les statistiques que l'on considère (X_G, X_T, X_A ...) ne sont en général pas indépendantes et de fait leur distribution sous $H0$ n'est pas forcément triviale à obtenir. Dans ce cas, ces distributions ne peuvent pas être approchée par une loi de probabilité connue et doivent être déterminée par une approche exacte ou empirique.

A titre d'exemple, nous avons illustré figure 2.2 la différence entre la distribution théorique de $X_{T+G} = X_T + X_G$ si les statistique X_T et X_G étaient indépendantes[3] et la distribution observée en pratique. Même si la différence n'est pas très prononcée, les deux distributions ne sont pas les mêmes. En particulier, si on réalise un test au niveau 5% sous l'hypothèse d'indépendance[4], on obtient en réalité un taux d'erreur de type-I réel de 8%. Cet exemple illustre assez bien la nécessité de réaliser avec précaution un test construit à partir d'une combinaison de statistiques ou de tests.

Problématiques liées aux analyses simple-marqueur

Une analyse simple-marqueur peut paraître au premier abord assez simple à mettre en place. Outre les considérations d'ordre statistique du type la taille de l'échantillon, le choix du test[5] et le choix de l'approche pour estimer la p-value[6], elles soulèvent un certains

[3] soit un χ^2 à 3 degrés de liberté

[4] soit 7.87 pour un χ^2 à 3 degrés de liberté

[5] test de score, de rapport de vraisemblance ou de Wald

[6] lié à la validité d'une approximation aymptotique dans certaines conditions et au temps d'exécution

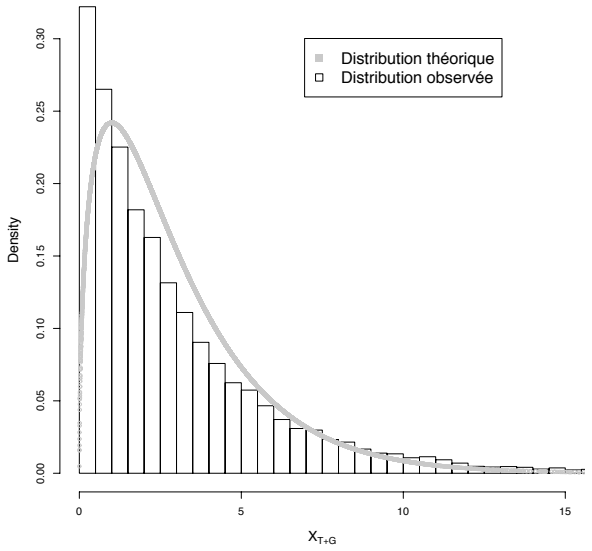


Fig. 2.2 – **Distribution observée de X_{T+G} :** comparaison avec la distribution théorique sous l'hypothèse d'indépendance entre X_T et X_G : $\chi^2(3)$. Test d'adéquation des deux distributions de Kolmogorov-Smirnov : $p\text{v} < 0.05$.

nombre d'autres problématiques en terme de stratégie d'analyse et nous en développons deux.

Choix des statistiques : il existe un certain nombre de statistiques et donc de tests d'association. Chacun repose sur des hypothèses statistiques et biologiques sensiblement différentes. A ces tests s'ajoutent toutes les combinaisons de statistiques et de tests possibles. Devant l'étendue de ce choix, il est raisonnable de se demander :

Quelle stratégie faut-il adopter ? Quel statistique ai-je intérêt à utiliser, dans quelle situation ? Est-il judicieux de réaliser plusieurs tests en même temps à travers une combinaison de statistiques ou de tests ? Tous les tests sont-ils réellement appropriés quel que soit le contexte biologique sous-jacent ?

Afin de répondre à cet ensemble de questions, nous avons mené une étude de puissance pour aider à proposer ce qui ne constitue peut-être pas pour nous la meilleure stratégie - à supposer qu'il en existe une - mais la plus éclairée et la plus adaptée. La section suivante traite de la mise en place de cette étude, du plan expérimental et des principaux résultats que nous avons obtenus tandis que les deux sections suivantes traitent en détails de cas qui nous semblent particuliers : le test allélique (p. 69) et le test d'Hardy-Weinberg (p. 77).

Choix du niveau de significativité : une autre thématique dont nous avons déjà parlé en introduction est le problème du test-multiple (p. 30). Pour rappel, il s'agit de l'inflation du taux d'erreur de type-I lorsque l'on réalise de façon indépendante un grand nombre de tests à un niveau fixé. Nous avons introduit deux quantités qui représentent chacune un critère alternatif de sélection : le FWER et le FDR. Dans ce chapitre, nous nous proposons de promouvoir une autre quantité apparue récement dans la littérature et qui semble particulièrement adaptée au problème de test-multiple en génomique à grande échelle : le FDR Local (p.87)

2.4 Étude de puissance

Une étude de puissance consiste à évaluer la puissance (π) d'une approche d'analyse statistique, c'est à dire sa capacité à rejeter l'hypothèse $H0$ quand celle-ci est fausse :

$$\pi(\alpha) = \mathbb{P}_{H1}(\mathcal{S} \geqslant t_\alpha).$$

Comme énoncé en introduction, la puissance dépend d'un certains nombre de facteurs dont le niveau α ainsi que la distribution de la statistique sous $H1$. Comme $H1$ représente tout ce qui n'est pas $H0$, il est impossible d'avoir une distribution de la statistique sous $H1$ en toute généralité et une étude de puissance se fera donc par l'intermédiaire d'un certain nombre de $H1$ différentes, représentatives de l'ensemble des alternatives possibles. Dans les cas de tests d'indépendance appliqués à des tables de contingence, l'hypothèse $H1$ se traduira par des paramètres du modèle d'échantillonnage s'écartant de ceux définis par l'hypothèse nulle. L'idée est de déterminer ces paramètres de façon à ce qu'ils représentent au mieux la réalité biologique de la maladie. Pour cela, nous introduisons un **modèle génétique**.

Une fois ces paramètres déterminés, l'objectif est d'obtenir la distribution de la statistique d'association que l'on considère sous cette alternative. Il existe plusieurs démarches mathématiques possibles : les deux approches les plus populaires en génétique reposent sur des simulations de Monte-Carlo pour l'une et sur l'approximation de la distribution sous $H1$ par un χ^2 décentré pour l'autre. A ces deux approches nous ajoutons la Delta-Méthode telle qu'elle est traditionnellement employée, ainsi qu'une extension de celle-ci permettant de traiter des tests plus complexes, construits par exemple à partir de méta-statistiques.

Enfin, et après avoir abordé le protocole expérimental utilisé pour cette étude, nous présenterons l'ensemble des résultats obtenus. Les cas des tests allélique et d'Hardy-Weinberg faisant l'objet de sections à part entière, nous nous limitons ici aux cas du test génotypique (X_G), du test de tendance (X_T) ainsi que de deux combinaisons de ces tests reposant sur les statistiques $X_{T+G} = X_T + X_G$ et $X_{T\times G} = X_T \times X_G$

Le modèle génétique

L'objectif est de déterminer les paramètres du modèle d'échantillonage. On considère les N_D cas et les N_H témoins comme étant échantillonnés indépendamment suivant une multinomiale $\mathcal{M}(N_D, p_{D_0}, p_{D_1}, p_{D_2})$ pour les cas et $\mathcal{M}(N_H, p_{H_0}, p_{H_1}, p_{H_2})$ pour les témoins. N_D et N_H étant fixés, il reste à déterminer les probabilités pour chaque individu de porter un génotype sachant qu'il appartient au groupe des cas (p_{D_i}) ou des témoins (p_{H_i}).

Soit un marqueur bi-allélique de type SNP par exemple ; soient p_0, p_1 et p_2 les proportions génotypiques du marqueur dans la population déterminées à partir de ses proportions alléliques (p_a et $p_A = 1 - p_a$) et du coefficient de consanguinité ($\mathcal{F}$) :

$$\left\{\begin{array}{lcl} p_0 & = & p_a^2 + \mathcal{F}p_a(1-p_a) \\ p_1 & = & 2p_a(1-p_a) - 2\mathcal{F}p_a(1-p_a) \\ p_2 & = & (1-p_a)^2 + \mathcal{F}p_a(1-p_a) \end{array}\right.$$

Soit K_p la prévalence de la maladie[7] dans une population donnée. Soient f_0, f_1 et f_2 les pénétrances[8] associées aux génotypes aa, aA et AA respectivement. A partir des pénétrances on peut définir n'importe quel mode de transmission de la maladie (noté MOI pour *mode of inheritance*). Si l'on considère A comme étant l'allèle de susceptibilité :

$$\left\{\begin{array}{llll} f_1 = f_0 & \text{et} & f_2 = f_0 + c & \text{définit le mode récessif (R).} \\ f_1 = f_0 + c & \text{et} & f_2 = f_1 + c = f_0 + 2c & \text{définit le mode additif (A).} \\ f_1 = f_0 \times c & \text{et} & f_2 = f_1 \times c = f_0 \times c^2 & \text{définit le mode multiplicatif (M).} \\ f_1 = f_0 + c & \text{et} & f_2 = f_1 & \text{définit le mode dominant (D).} \\ f_2 = f_0 & \text{et} & f_1 = f_0 + c & \text{définit le mode sur-dominant (S).} \end{array}\right.$$

En notant que $K_p = p_0 f_0 + p_1 f_1 + p_2 f_2$, l'on peut à partir de f_2 ou f_1, de c, du mode de transmission, de K_p, de p_a ou p_A, de $\mathcal{F}$ et des formules de Bayes, déterminer les paramètres recherchés :

$$\begin{array}{rcl} (p_{D_0}, p_{D_1}, p_{D_2}) & = & \left(\dfrac{f_0 p_0}{K_p}, \dfrac{f_1 p_1}{K_p}, \dfrac{f_2 p_2}{K_p}\right), \\ (p_{H_0}, p_{H_1}, p_{H_2}) & = & \left(\dfrac{(1-f_0)p_0}{1-K_p}, \dfrac{(1-f_1)p_1}{1-K_p}, \dfrac{(1-f_2)p_2}{1-K_p}\right). \end{array}$$

Une façon de gagner un paramètre consiste à définir le modèle en fonction des risques relatifs plutôt que des pénétrances. Soit $RR_i = \frac{f_i}{f_0}$:

$$\left\{\begin{array}{ll} RR_1 = 1 & \rightarrow \text{R} \\ RR_1 = \frac{RR_2+1}{2} & \rightarrow \text{A} \\ RR_1 = \sqrt{RR_2} & \rightarrow \text{M} \\ RR_1 = RR_2 & \rightarrow \text{D} \\ RR_2 = 1 & \rightarrow \text{S} \end{array}\right.$$

[7] c'est à dire la probabilité de développer la maladie

[8] c'est à dire les risques pour un individu de contracter la maladie sachant qu'il possède tel ou tel génotype

Le modèle génétique met finalement en jeu 5 paramètres à fixer (p_a ou p_A, $\mathcal{F}$, K_p, RR_2 ou RR_1, MOI), et dans ce contexte, $H0 : \{RR_1 = RR_2 = 1\}$. Ce modèle permet de simuler des situations où le SNP considéré est le SNP étiologique (association directe). L'on peut aussi mettre en place des situations où le SNP que l'on considère est un marqueur en déséquilibre de liaison avec le SNP étiologique (association indirecte). Dans ce cas il est nécessaire de prendre en compte le déséquilibre de liaison et d'introduire dans notre modèle un paramètre correspondant. Nous avons choisi le coefficient de corrélation (r^2) présenté en introduction (p. 9)

Méthodes d'approximation de la distribution sous $H1$

Nous introduisons ici quatre méthodes d'approximation de la distribution d'une statistique d'association sous $H1$: les trois les plus utilisées dans la littérature à savoir les simulations de Monte-Carlo, le χ^2 décentré et la Delta-Méthode auxquelles nous ajoutons une extension de la Delta-Méthode à l'ordre 2, appelée dans ce texte Forme Quadratique.

- Monte-Carlo : cette technique est tout à fait similaire à ce que nous avons vu pour estimer la p-value. Le but est de générer B tables de contingences $(\mathcal{T}^{(1)}, ..., \mathcal{T}^{(B)})$ à partir du modèle d'échantillonnage afin d'avoir une distribution empirique de la statistique $(\mathcal{S}^{(1)}, ..., \mathcal{S}^{(B)})$ sous le modèle $H1$ considéré. A partir de cette distribution l'on déduit une estimation de la puissance par :

$$\widehat{\pi}(\alpha) = \frac{\sharp\{\mathcal{S}^{(i)} \geqslant t_\alpha\}}{B},$$

où $\sharp$ est l'opérateur cardinal. Comme pour le calcul de la p-value, cette approche est dans le cas de la puissance facile à mettre en place et donc très largement utilisée, en particulier dans le champ de la Statistique appliquée à la Génétique où la distribution sous $H1$, comme sous $H0$, peut être difficile voir impossible à approcher analytiquement (Longmate 2001). Néanmoins ce type d'approche requiert un certain temps d'exécution dont dépend la précision de l'estimation. De façon analogue à la p-value, $\widehat{\pi} \times B$ est distribué suivant une binomiale $\mathcal{B}(B, \pi)$ et $\widehat{\pi}$ peut être approché par une distribution normale $\mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{B}\right)$. Cela nous permet d'obtenir pour $\widehat{\pi}$ l'intervalle de confiance à 95% qui diminue avec une vitesse de l'ordre de $1/\sqrt{B}$:

$$\left[\widehat{\pi} - 1.96\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{B}}; \widehat{\pi} + 1.96\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{B}}\right].$$

- χ^2 décentré : En 1958, Mitra démontre que sous l'hypothèse alternative, une statistique de χ^2 suit asymptotiquement une distribution du χ^2 décentrée $\chi^2(k, \lambda)$ où k et le degré de liberté et λ le paramètre de non-centralité dont l'expression dépend de la

statistique considérée. Ce dernier sera indiqué dans le protocole expérimental pour les statistiques que nous considérons). Une fois λ déterminé[9], l'on peut déduire une estimation asymptotique de la puissance :

$$\pi(\alpha) = \mathbb{P}(\chi^2_{1-t_\alpha}(k,\lambda) \geqslant t_\alpha).$$

Il s'agit d'une approche asymptotique qui est de fait restreinte à certaines conditions d'application[10]. Elle reste néanmoins une approche rapide, considérée avec un intérêt croissant dans les études d'association où les statistiques suivent bien souvent des distributions du χ^2 (Sham et al 2000, Gordon et al 2002, Kang et al 2004).

- Delta-Méthode : elle est utilisée pour approcher la distribution de $\mathcal{S}$ sous $H1$ à partir de la distribution des valeurs prises par la table de contingence $X = (D_0, D_1, D_2, H_0, H_1, H_2)$, en faisant l'hypothèse que celles-ci suivent une distribution multinomiale et peuvent donc être approchées par une distribution normale $\mathcal{N}(M, \Sigma)$ où M est l'espérance de X et Σ la matrice de variance-covariance données par :

$$M = \begin{pmatrix} N_D \times p_{D_0} \\ N_D \times p_{D_1} \\ N_D \times p_{D_2} \\ N_H \times p_{H_0} \\ N_H \times p_{H_1} \\ N_H \times p_{H_2} \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} N_D p_{D_0}(1-p_{D_0}) & -N_D p_{D0} p_{D_1} & -N_D p_{D_0} p_{D_2} & 0 & 0 & 0 \\ -N_D p_{D_0} p_{D_1} & N_D p_{D_1}(1-p_{D_1}) & -N_D p_{D_1} p_{D_2} & 0 & 0 & 0 \\ -N_D p_{D_0} p_{D_2} & -N_D p_{D_1} p_{D_2} & N_D p_{D_2}(1-p_{D_2}) & 0 & 0 & 0 \\ 0 & 0 & 0 & N_H p_{H_0}(1-p_{H_0}) & -N_H p_{H_0} p_{H_1} & -N_H p_{H_0} p_{H_2} \\ 0 & 0 & 0 & -N_H p_{H_0} p_{H_1} & N_H p_{H_1}(1-p_{H_1}) & -N_H p_{H_1} p_{H_2} \\ 0 & 0 & 0 & -N_H p_{H_0} p_{H_2} & -N_H p_{H_1} p_{H_2} & N_H p_{H_2}(1-p_{H_2}) \end{pmatrix}$$

En appliquant un développement de Taylor à l'ordre 1 de $\mathcal{S} = f(X)$ autour M, on peut approcher $\mathcal{S}$ par :

$$\mathcal{S} \approx f(M) +^t (X - M) \times \nabla f(M),$$

où ^{t} est l'opérateur de transposition et ∇f le gradient de f. Ce développement à l'ordre 1 permet d'approcher la distribution de $\mathcal{S}$ par une distribution normale $\mathcal{N}(m, \sigma)$ avec $m = f(M)$ et $\sigma = \sqrt{^t\nabla f(M) \times \Sigma \times \nabla f(M)}$. A partir de cette approximation il est possible d'obtenir une estimation de la puissance :

$$\pi(\alpha) = 1 - \Phi\left(\frac{t_\alpha - m}{\sigma}\right),$$

[9] ce qui peut se faire en remarquant que $\lambda = \mathbb{E}(\chi^2(k,\lambda)) - 2k$

[10] à savoir les mêmes qui doivent être appliquées à l'estimation de la p-value par l'approximation asymptotique du χ^2

où Φ est la fonction de répartition[11] d'une normale centrée réduite $\mathcal{N}(0,1)$. La qualité de cette approche asymptotique et numérique dépend de la qualité de l'approximation normale faite sur X et sur $\mathcal{S}$.

- Forme Quadratique : pour les cas où l'approximation de la distribution de $\mathcal{S}$ par une normale sous $H1$ n'est pas réaliste, nous proposons d'étendre le développement de Taylor effectué précédemment à l'ordre 2 ; on obtient alors une approximation plus précise reposant sur la distribution d'une forme quadratique en variables normales[12] :

$$\mathcal{S} \approx f(M) +^t (X - M) \times \nabla f(M) + \frac{1}{2}\, ^t(X - M) \times \nabla^2 f(M) \times (X - M),$$

où $\nabla^2 f$ est la Hessienne de f. Dans le cas précédent, la calcul de la puissance nécessitait la CDF d'une distribution normale. Ici, la distribution de $\mathcal{S}$ est approchée par une combinaison de χ^2 dont la CDF est beaucoup moins évidente à obtenir. Les détails techniques à ce sujet se trouvent en annexe page 189.

Protocole expérimental

Statistiques considérées : nous considérons quatre statistiques : X_G et X_T correspondantes aux tests génotypique et de tendance respectivement, en ce qui concerne les statistiques classiques, et $X_{T+G} = X_T + X_G$ et $X_{T\times G} = X_T \times X_G$ en ce qui concerne les méta-statistiques. Afin de permettre l'estimation de la puissance par χ^2 décentré, nous fournissons par ailleurs les valeurs des paramètres de décentralisation λ_G et λ_T en fonction des paramètres du modèle génétique et correspondantes à X_G et X_T respectivement :

$$\lambda_G = N_D N_H \times \sum_0^2 \frac{\left(\frac{f_i p_i}{K_p} - \frac{(1-f_i)p_i}{1-K_p}\right)^2}{N_D \frac{f_i p_i}{K_p} + N_H \frac{(1-f_i)p_i}{1-K_p}}$$

$$\lambda_T = N_D N_H \times \frac{\left[\sum_1^2 i \left(\frac{(1-f_i)p_i}{1-K_p} - \frac{f_i p_i}{K_p}\right)\right]^2}{\sum_1^2 i \left(N_D \frac{f_i p_i}{K_p} + N_H \frac{(1-f_i)p_i}{1-K_p}\right) - \frac{\left[\sum_1^2 i \left(N_D \frac{f_i p_i}{K_p} + N_H \frac{(1-f_i)p_i}{1-K_p}\right)\right]^2}{N}}$$

Paramètres fixés pour toute l'étude : $K_p = 0.05$, $B = 5,000$ et $N_D = N_H = 500$.

Paramètres à faire varier : MOI = {R, A, M, D, S}, $p_A = [0,1]$, $r^2 = [0,1]$, $\mathcal{F} = [-1,1]$ et $RR_2 = \{1.5, 2\}$ pour les modes de transmission récessif, additif, multiplicatif et

[11] CDF pour *Cumulative Distribution Function*
[12] QFNV pour *Quadratic Form in Normal Variables*

dominant et $RR_1 = \{1.5, 2\}$ pour le mode de transmission sur-dominant. Par ailleurs, pour le jeu de paramètres choisi, les résultats concernant les modes de transmission additifs et multiplicatifs sont très similaires. Par conséquent, dans un souci de simplicité, nous ne présentons pas les résultats obtenus à partir du mode de transmission multiplicatif.

Résultats

- Comparaison des méthodes d'approximation de $H1$: l'objectif de ce paragraphe est de comparer les différentes méthodes d'estimation de la puissance que nous avons introduites. Notre référence sera l'estimation obtenue par Monte-Carlo qui au bout d'un certain nombre de simulations se rapproche avec précision de la vraie valeur de π[13].

La figure 2.3 page 64 illustre les résultats concernant les statistiques simples (X_T et X_G). L'approche χ^2 décentré est de façon non surprenante totalement adaptée aux statistiques qui suivent asymptotiquement une distribution du χ^2 et fournit donc des estimations très précises. Comme la distribution d'un χ^2 décentré est un cas particulier d'une Forme Quadratique, celle-ci donne également de très bon résultats. En contrepartie la Delta-Méthode sous-estime la puissance dans les deux cas ce qui met en évidence le fait que l'approximation de la distribution des statistiques considérées sous $H1$ par une distribution normale n'est pas du tout réaliste. On peut néanmoins noter que cette approche apporte de meilleurs estimations pour le test de tendance que pour le test génotypique ; cette différence de qualité d'estimation peut être due au fait que la distribution de X_T sous $H1$ serait moins éloigné d'une distribution normale que celle de X_G ; le test de normalité de Shapiro-Wilk semble aller dans ce sens : la valeur moyenne de la statistique correspondante est de 0.69 pour un échantillon tiré suivant un $\chi^2(1)$ et de 0.81 pour un échantillon tiré suivant un $\chi^2(2)$[14]. Quoi qu'il en soit, il n'est pas raisonnable d'approcher les distributions de X_T et X_G par des distributions normales et l'utilisation de la Delta-Méthode à l'ordre 1 n'apparaît donc pas comme une bonne solution.

Par ailleurs il a été suggéré dans la littérature que des facteurs tels que le rapport cas-témoins (N_D/N_H), la fréquence allélique (p_a) et la taille de l'échantillon (N) pouvaient avoir un impact sur la précision des approches analytiques par rapport à du Monte-Carlo (Ji et al 2005). Nous avons cherché à mettre en évidence de tels effets en faisant varier ces paramètres ($N_D/N_H = \{0.04; 0.2; 1\}$ et $N = \{40; 200; 1000\}$). Nous n'aboutissons cependant pas aux mêmes constatations et ne développons donc pas plus la question.

La figure 2.4 page 65 donne les résultats concernant l'estimation de la puissance pour les méta-statistiques (X_{T+G} et $X_{T\times G}$). Dans ce cas l'approche χ^2 décentré n'est pas applicable puisque les distributions résultantes de X_{T+G} et de $X_{T\times G}$ ne suivent asymptotiquement pas une distribution du χ^2 de fait de la non-indépendance de X_G avec X_T.

[13] à l'exception des valeurs très proches de 0 ou de 1 qui demandent un nombre conséquent de simulations, mais nous ne nous plaçons pas dans ces cas

[14] ce qui est contraire à ce que l'on pourrait attendre avec des distributions de $\chi^2(1)$ et $\chi^2(2)$

La Delta-Méthode continue de donner des estimations aussi peu précises. En revanche la Forme Quadratique apporte de très bonnes estimations en ce qui concerne X_{T+G} ; comme souligné précédemment, une combinaison linéaire de statistiques distribuées suivant un χ^2 indépendants ou non est une Forme Quadratique ce qui explique que cette méthode traite si bien X_{T+G}. Pour ce qui est de $X_{T\times G}$, ni l'ordre 1 ni l'ordre 2 de la Delta-Méthode ne donnent de résultats satisfaisants ; on peut penser que le produit de χ^2 requiert un ordre supérieur, pour lequel la détermination de la CDF serait numériquement très coûteuse voir irréalisable en pratique.

- Puissance des statistiques simples : les résultats sur les statistiques simples sont relativement évidents (figures 2.5, 2.6 et 2.7 p. 66, 67 et 68). Le test de tendance est efficace pour tester les cas pour lesquels il a été construit ; en l'occurrence, il est le plus puissant pour traiter les modes de transmission additif et multiplicatif. En revanche il est très mauvais pour traiter un mode de transmission sur-dominant, conceptuellement le plus éloigné du mode additif. Le test génotypique présente les meilleurs puissances pour trois des quatre modes de transmission illustrés (R, D et S). Ce n'est pas surprenant puisque ce test est très général et ne fait aucune hypothèse sur la façon dont les proportions génotypiques dévient de l'hypothèse nulle. Par ailleurs même pour les modes de transmission pour lesquels il est le plus faible (A et M), il talonne de près le test de tendance.

- Puissance des méta-statistiques : de façon générale les puissances de X_{T+G} et $X_{T\times G}$ s'intercalent entre celles de X_T et X_G (figures 2.5, 2.6 et 2.7 p. 66, 67 et 68). Nous avons pu constater des résultats de la même nature pour des combinaisons de tests du type *on rejette* $H0$ *si le test de tendance et/ou le test génotypique rejettent* $H0$ (résultats non présentés). Il s'avère ici que la combinaison des deux statistiques à travers une simple somme présente de bien meilleurs résultats que le produit. On observe également dans la littérature que les méta-statistiques fondées sur des sommes semblent être plus performantes que celles fondées sur des produits (Hoh et al 2001, Song et Elston 2006). Enfin, l'on peut noter que si les méta-statistiques ne présentent pas les meilleures puissances, elles se trouvent cependant proches du meilleur test pour la plupart des situations considérées.

- Discussion : l'étude de puissance est un outil important en statistique pour comparer l'efficacité de différentes méthodes d'analyse ou pour aider à mettre en place une étude. Avec l'accumulation de nouvelles stratégies d'analyse qui accompagne aujourd'hui l'accumulation de grandes quantités de données, L'Épidémiologie Génétique ne peut faire l'économie de telles études. Nous avons considéré ici le problème de l'estimation de la puissance dans le cadre des analyses simple-marqueur.

L'approche par Monte-Carlo, facile à mettre en place, est souvent celle employée. Elle demande néanmoins un certain temps d'exécution et la précision des estimations est directement liée au nombre de simulations réalisées (B). En particulier, la largeur de

l'intervalle de confiance de l'estimation diminue avec une vitesse $1/\sqrt{B}$. L'approche χ^2 décentrée est naturellement adaptée à toute statistique qui suit un χ^2. La Delta-Méthode à l'ordre 1 fait l'hypothèse d'une distribution normale de la statistique sous $H1$ ce qui n'est pas le cas des statistiques que nous avons considérées. Cependant dans la littérature, cette approche est utilisée avec succès pour fournir des estimations précises de puissances du test allélique et de tendance (Slager et Schaid 2001, Jackson et al 2002). Brièvement, la distribution normale que nécessite l'application de cette méthode peut être obtenue en considérant par exemple la statistique : $z_T \sim \mathcal{N}(0,1)$ sous $H0$ avec $z_T^2 = X_T$ au lieu de X_T directement. Dans ce cadre la Delta-Méthode donne de très bonnes estimations et revient en fait à une simplification analytique de la Forme Quadratique ; cependant son application est restreinte à des statistiques de type z-score ou par extension à des statistiques distribuées suivant un $\chi^2(1)$, ce qui rend de fait cette approche moins générale que celle fondée sur le χ^2 décentré. A ces trois méthodes d'estimation, nous avons ajouté une approche fondée sur la Delta-Méthode développée à l'ordre 2 : Forme Quadratique. Cette approche produit de bons résultats et peut traiter aussi bien des statistiques simples (X_G, X_T...) que des combinaisons linéaires de ces mêmes statistiques ce qui représente un avantage sur les autres approches. Malgré une évaluation de la CDF bien moins évidente, elle représente une alternative moins coûteuse en temps d'exécution que le Monte-Carlo, plus générale que le χ^2 décentré et plus précise que la Delta-Méthode à l'ordre 1.

Les résultats sur les statistiques simples étaient prévisibles : le test de tendance est le meilleur dans des cas additifs ou multiplicatifs tandis que le test génotypique, plus général, présente les meilleures puissances dans les cas récessifs, dominants et sur-dominant sans par ailleurs montrer de grosses pertes de puissance dans les situations intermédiaires.

A première vue, la stratégie qui consiste à combiner des statistiques ou des tests avec comme objectif d'obtenir plus de puissance échoue. Dans la majeure partie des cas, les puissances résultantes s'intercalent entre celles des tests simples. S'ils ne sont pas les tests les plus efficaces quelle que soit la situation considérée, ils ne sont par ailleurs pas non plus ceux montrant la plus mauvaise puissance. De telles combinaisons permettent donc de mettre en place des stratégies d'analyse intermédiaires en terme de puissance, quel que soit le mode de transmission sous-jacent ce qui peux susciter de l'intérêt ; cependant l'estimation de quantités statistiques telles que la p-values ou la puissance s'avère bien moins évidente, ce qui peut mener à des résultats erronés lorsqu'elles ne sont pas réalisées avec précaution.

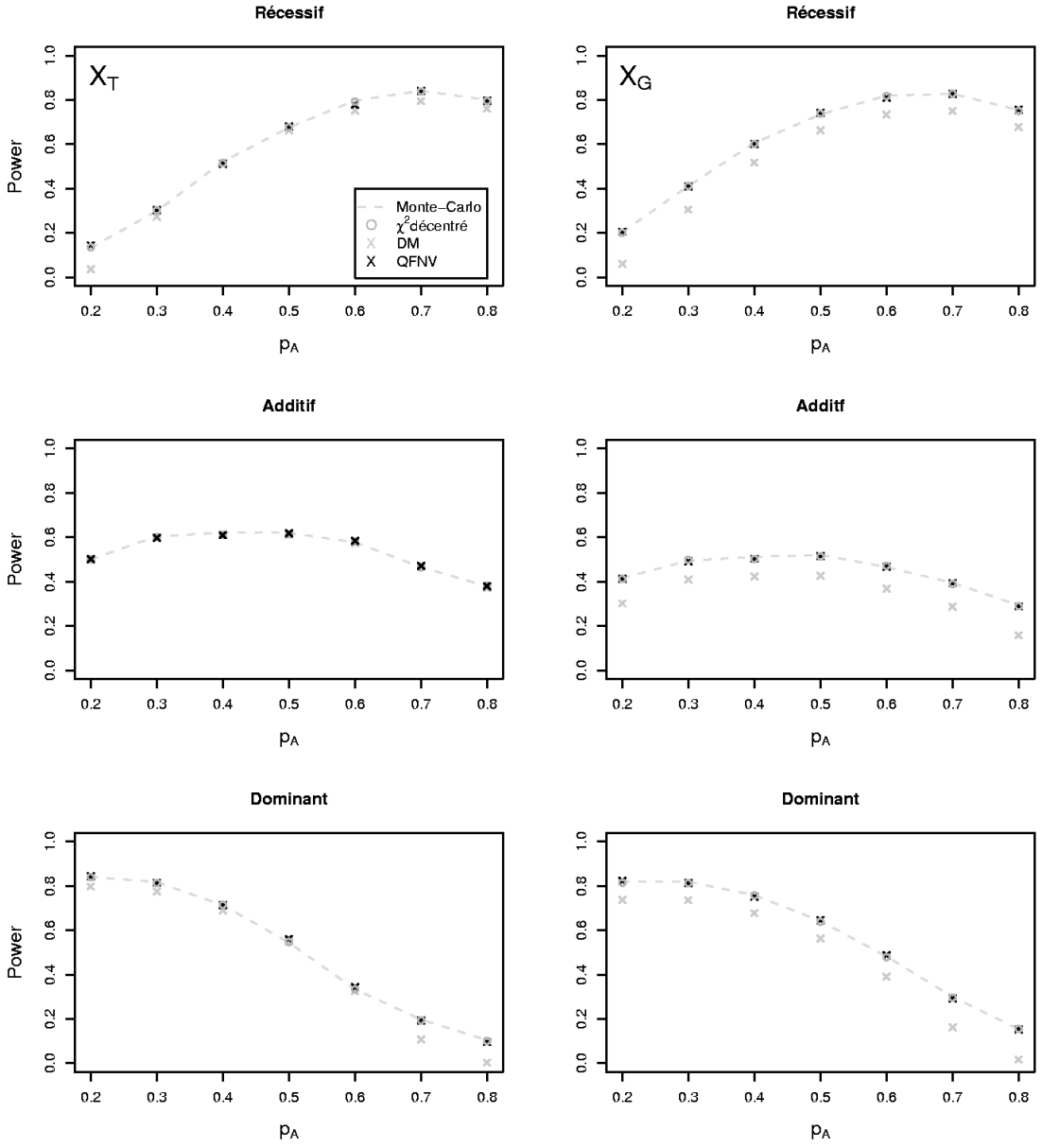


Fig. 2.3 – **Comparaison des méthodes d'estimation de puissance :** Monte-Carlo, χ^2 décentré, Delta-Méthode (DM) et Forme Quadratique (QFNV) appliquées aux statistiques simples du test de tendance (X_T) et génotypique (X_G) en fonction de la proportion allélique (p_A) au niveau 5% .

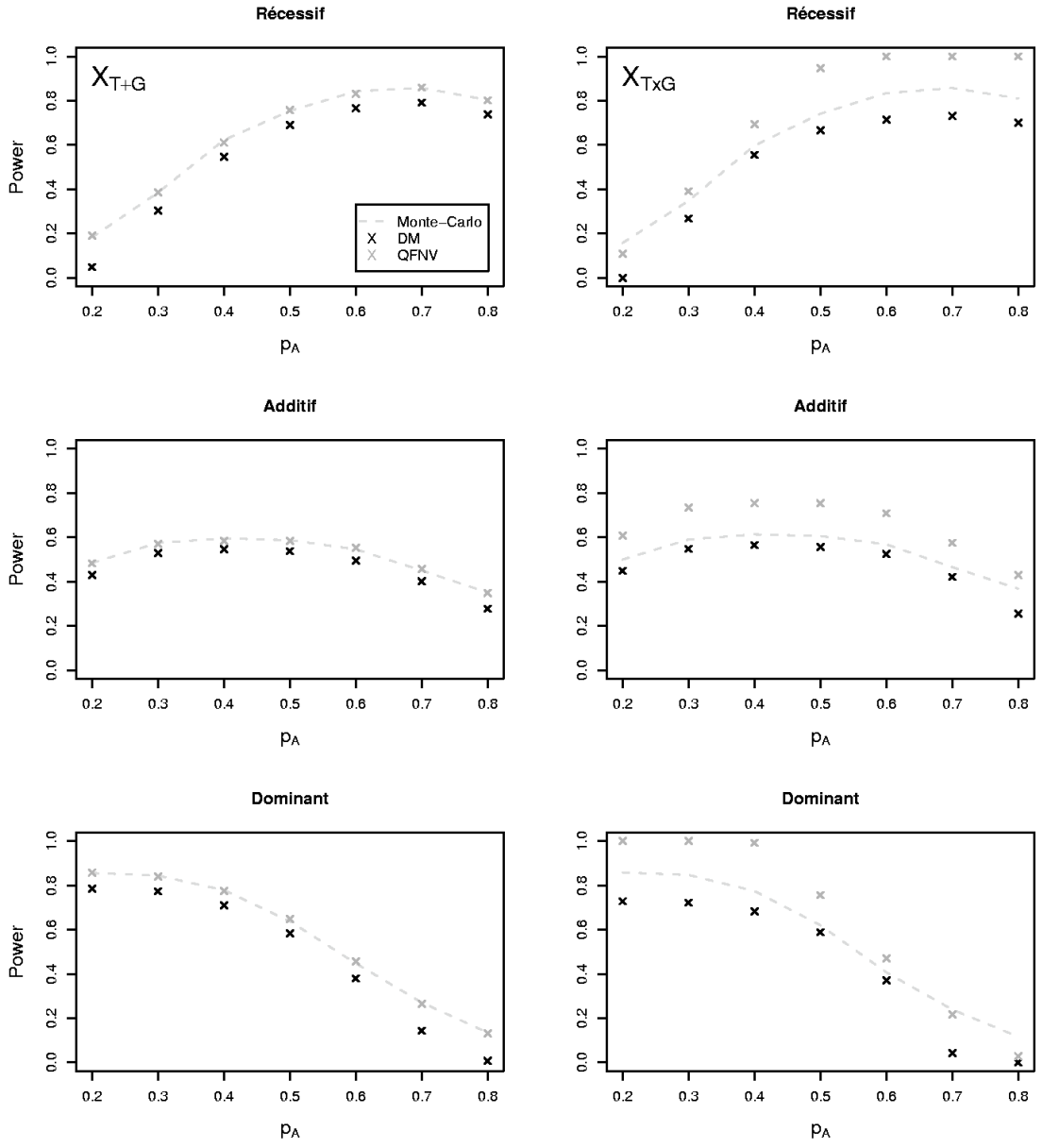


FIG. 2.4 – **Comparaison des méthodes d'estimation de puissance :** Monte-Carlo, χ^2 décentré, Delta-Méthode (DM) et Forme Quadratique (QFNV) appliquées aux méta-statistiques (X_{T+G} et $X_{T\times G}$) en fonction de la proportion allélique (p_A) au niveau 5%.

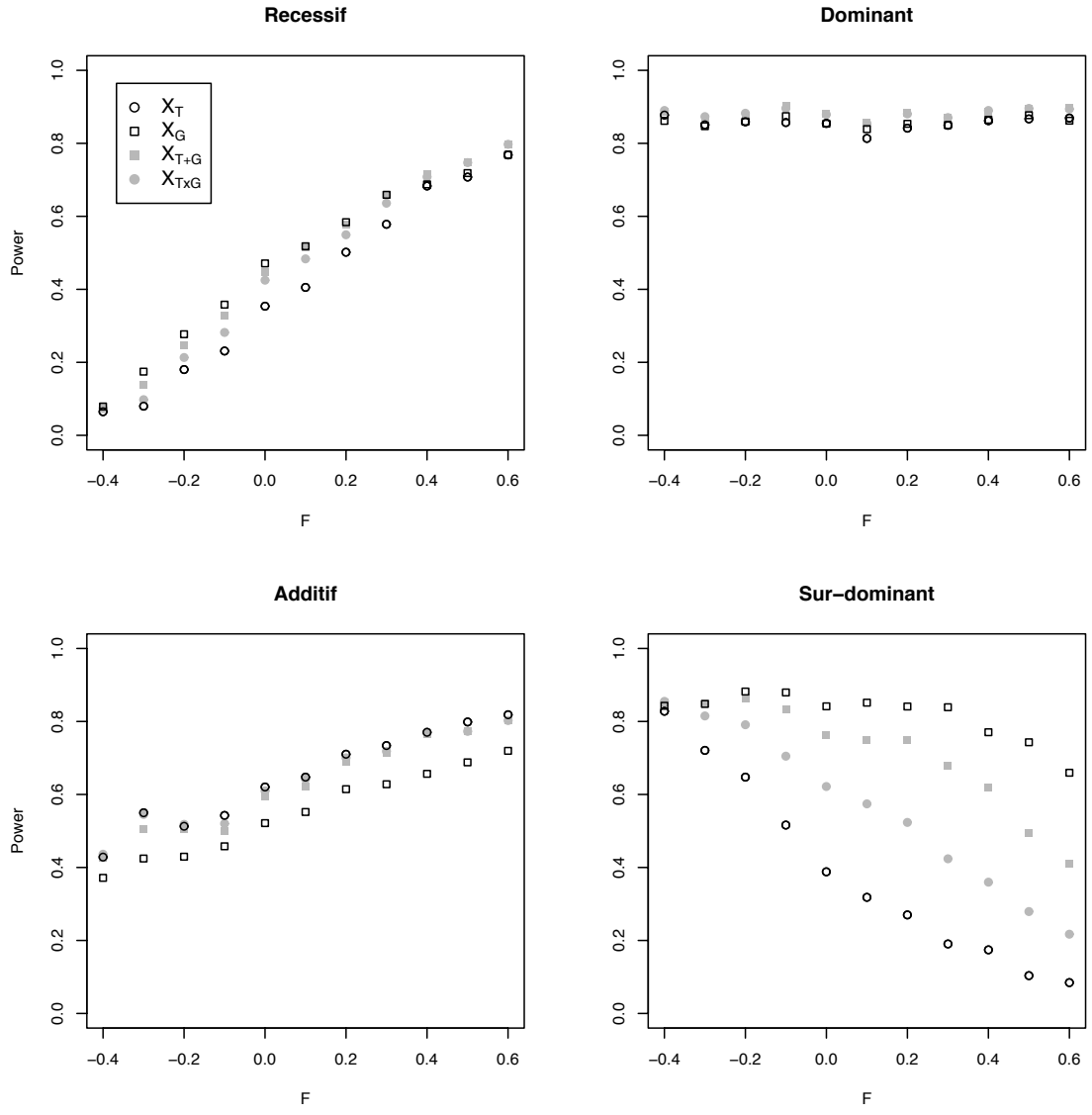


Fig. 2.5 – **Puissances de 4 tests d'association :** X_T, X_G, X_{T+G} et $X_{T\times G}$ en fonction du coefficient de consanguinité $\mathcal{F}$ au niveau 5% ($K_p = 0.05$, $p_A = 0.3$, $RR_2 = 1.5$ pour les modes de transmission récessif, additif et dominant ou $RR_1 = 1.5$ pour le mode de transmission sur-dominant, $N_D = N_H = 500$).

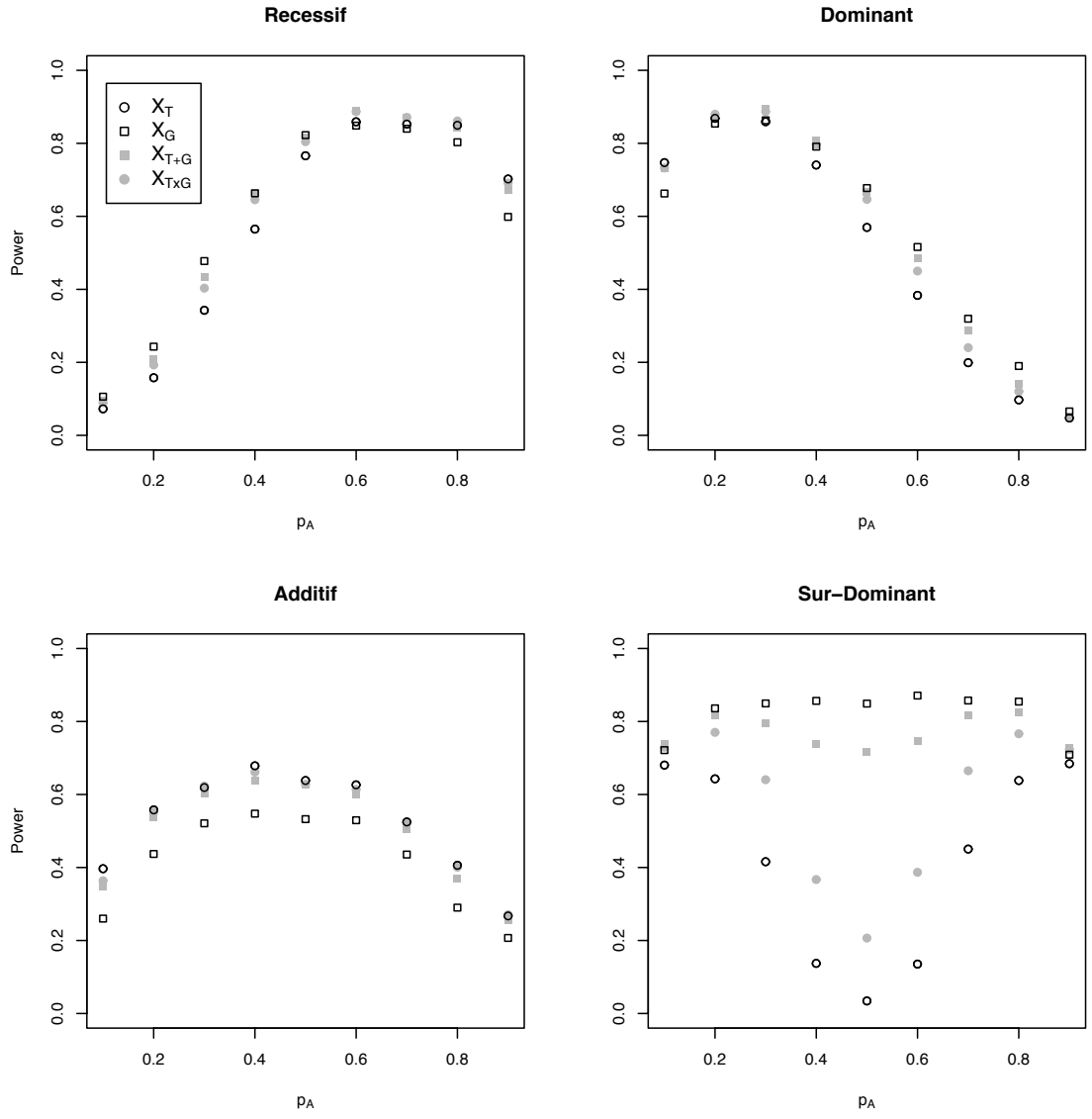


FIG. 2.6 – **Puissances de 4 tests d'association :** X_T, X_G, X_{T+G} et $X_{T\times G}$ en fonction des proportions alléliques p_A au niveau 5% ($K_p = 0.05$, $\mathcal{F} = 0$, $RR_2 = 1.5$ pour les modes de transmission récessif, additif et dominant ou $RR_1 = 1.5$ pour le mode de transmission sur-dominant, $N_D = N_H = 500$).

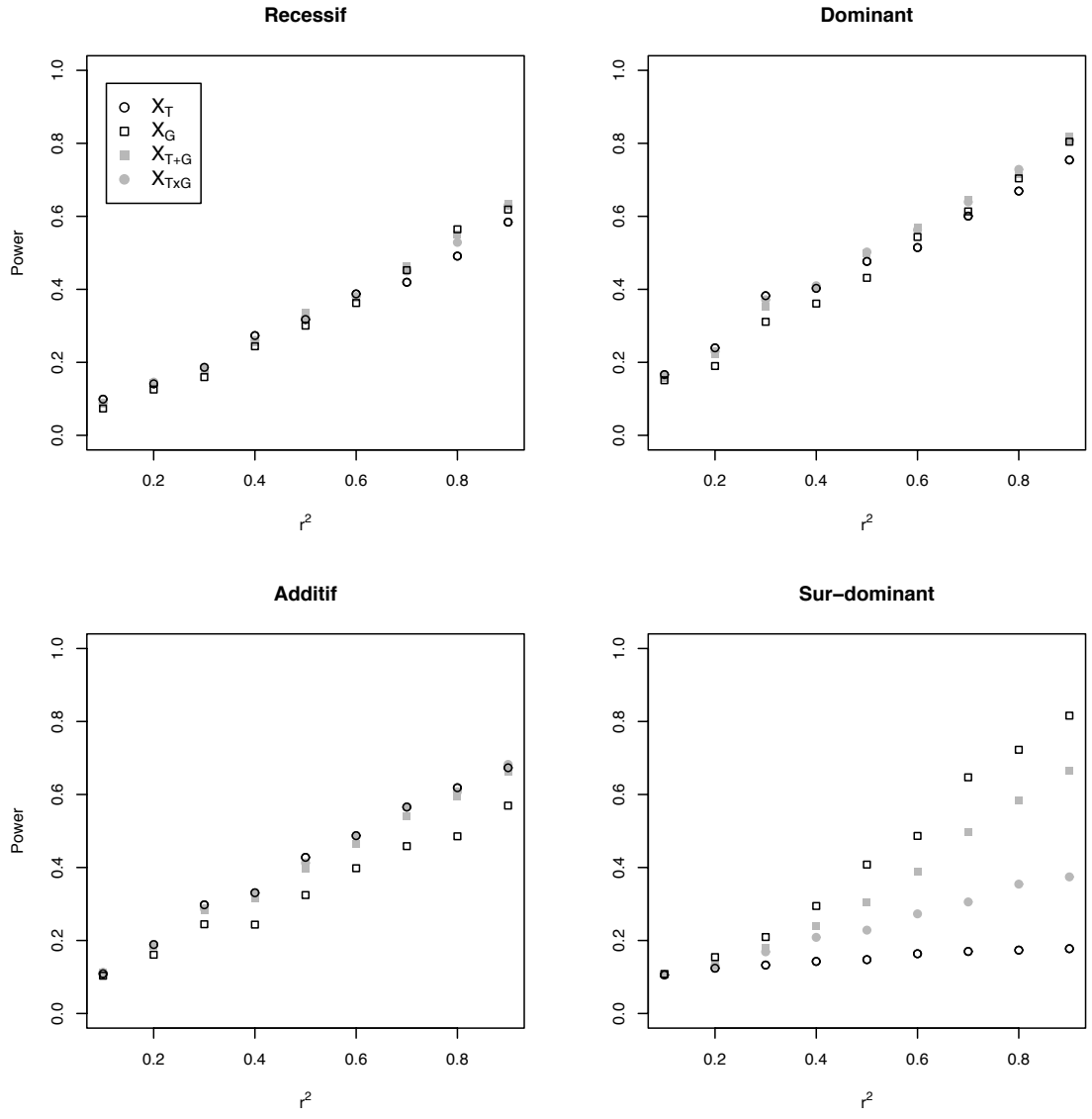


Fig. 2.7 – **Puissances de 4 tests d'association :** X_T, X_G, X_{T+G} et $X_{T\times G}$ en fonction du déséquilibre de liaison r^2 au niveau 5% ($K_p = 0.05$, $p_A = 0.4$, $q_A = 0.4$, $\mathcal{F} = 0$, $RR_2 = 1.5$ pour les modes de transmission récessif, additif et dominant ou $RR_1 = 1.5$ pour le mode de transmission sur-dominant, $N_D = N_H = 500$).

2.5 Cas particulier du test allélique

Le test allélique biaisé

La validité du test allélique a été récemment discutée et certains auteurs ne recommandent pas son utilisation (Sasieni 1997, Jackson et al 2002). Pour rappel, l'hypothèse nulle postule que la distribution des allèles est indépendante du statut cas ou témoin, mais aussi que les allèles sont échantillonnés indépendamment. Si les génotypes le sont, il est clair cependant que ce n'est pas le cas des allèles qui sont tirés par deux ; dans la table de contingence allélique, chaque individu contribue alors à deux observations. Par conséquent, la façon dont les allèles sont appariés[15], dévie X_A de la distribution supposée sous $H0$ et biaise l'estimation de la p-value. Au lieu du test allélique, Sasieni (1997) recommande d'utiliser le test de tendance dont l'auteur montre que la statistique (X_T) est exactement équivalente à X_A lorsque la population combinée des cas et des témoins observée est exactement à l'équilibre d'Hardy-Weinberg. Nous montrons en annexe (p. 191) que les deux statistiques sont asymptotiquement équivalente lorsque la population combinée est supposée à l'équilibre.

Le test allélique non-biaisé

Comme alternative plus naturelle au test allélique biaisé, nous proposons d'utiliser un test allélique non-biaisé, fondé sur la même statistique de Pearson (X_A) mais dont la distribution est déterminée en prenant en compte l'éventuelle dépendance entre les allèles. Slager et Jacobsen (1999) ont été les premiers à constater que le taux d'erreur de type-I du test allélique biaisé est altéré d'une manière prévisible et proposent une correction de X_A de sorte que la statistique suivent effectivement sous $H0$ une distribution du $\chi^2(1)$. Notre solution se fonde sur un calcul exact de la p-value passant par l'énumération des tables génotypiques possibles.

Plus précisément, le test allélique exact biaisé prend en compte toutes les tables 2×2 dont les marges correspondent à celles de la table allélique observée. Ainsi, la p-value correspondante ($p\mathrm{v}_b$ pour *biased*) est obtenue en sommant les probabilités des tables 2×2 $(\mathcal{T}_A^{(i)})$ pour lesquelles X_A est supérieure à la valeur observée X_A^{obs} :

$$p\mathrm{v}_b = \sum_{i|X_A^{(i)} \geqslant X_A^{\mathrm{obs}}} \mathbb{P}_{H0}(\mathcal{T}_A^{(i)}).$$

Le test allélique exact non-biaisé que nous proposons prend lui en compte toutes les tables 2×3 dont les marges correspondent à celles de la table génotypique observée. Ainsi, la

[15] où de façon équivalente la façon dont les proportions génotypiques dévient de l'équilibre d'Hardy-Weinberg

p-value correspondante ($p\text{v}_u$ pour *unbiased*) est obtenue en sommant les probabilités des tables 2×3 ($\mathcal{T}_G^{(i)}$) pour lesquelles X_A est supérieure à la valeur observée X_A^{obs} :

$$p\text{v}_u = \sum_{i|X_A^{(i)} \geqslant X_A^{\text{obs}}} \mathbb{P}_{H0}(\mathcal{T}_G^{(i)}).$$

De cette manière, les allèles ne sont plus considérés indépendamment et l'hypothèse $H0$ implique alors uniquement l'indépendance entre les allèles et la maladie. On peut par ailleurs noter que : **(i)** on se place ici dans un cadre de test exact conditionnel ; **(ii)** le test exact de Fisher appliqué à la table allélique conduit à une p-value biaisée alors que celui appliqué à la table génotypique correspond à un test exact génotypique et non plus allélique ; **(iii)** la p-value non-biaisée peut également être obtenue simplement par Monte-Carlo en permutant les labels cas-témoins.

Software

Le test allélique exact non-biasé nommé `fueatest`[16] (pour *Fast Exact and Unbiased Allelic Test*) est disponible sous différentes versions : R, C et Perl. Une attention particulière a notamment été portée à l'exécution rapide de ce test. La description précise de l'implémentation optimisée et réalisée par Karl Forner sort du cadre de cette thèse ; le lecteur intéressé pourra cependant se référer à l'article publié dans *Human Heredity* (2006) pour plus de détails. Brièvement les optimisations sont fondées sur une réduction sensible du nombre de tables à énumérer ainsi qu'une relation récursive entre les probabilités de chaque table $\mathbb{P}(\mathcal{T}_G^{(i)})$.

Afin d'illustrer le gain de temps réalisé par notre implémentation nous avons comparé le temps d'exécution avec celui d'une implémentation basique passant par l'énumération de toutes les tables. La comparaison s'est faite sur la base de 10,000 marqueurs avec une taille d'échantillon allant jusqu'à 8,000 individus (figure 2.8 p. 71). Bien que le temps d'exécution soit directement lié à la puissance de la machine, on constate un net avantage de notre implémentation optimisée ; pour de larges échantillons, le gain de temps peut être multiplié jusqu'à 6× et 12× pour des échantillons de 4,000 et 8,000 individus respectivement.

Application

- Données : nous avons appliqué les tests alléliques biaisé et non-biaisé sur les données AIM-Scan *genome-wide* concernant la sclérose en plaque. Le jeu de données comprend 66,990 SNPs ; l'ADN de 279 patients et 301 témoins suédois ont été génotypés en utilisant une puce *Affymetrix* 100K. La méthode utilisée pour déterminer les génotypes est celle décrite par Matsuzaki et al (2004).

[16] `http://stat.genopole.cnrs.fr/fueatest`

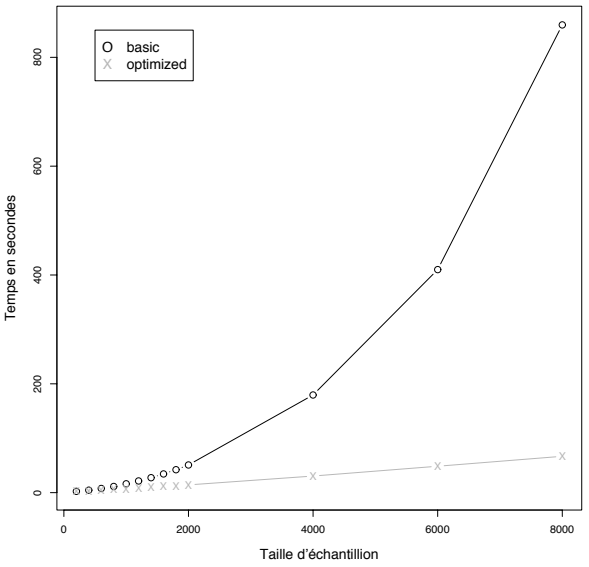
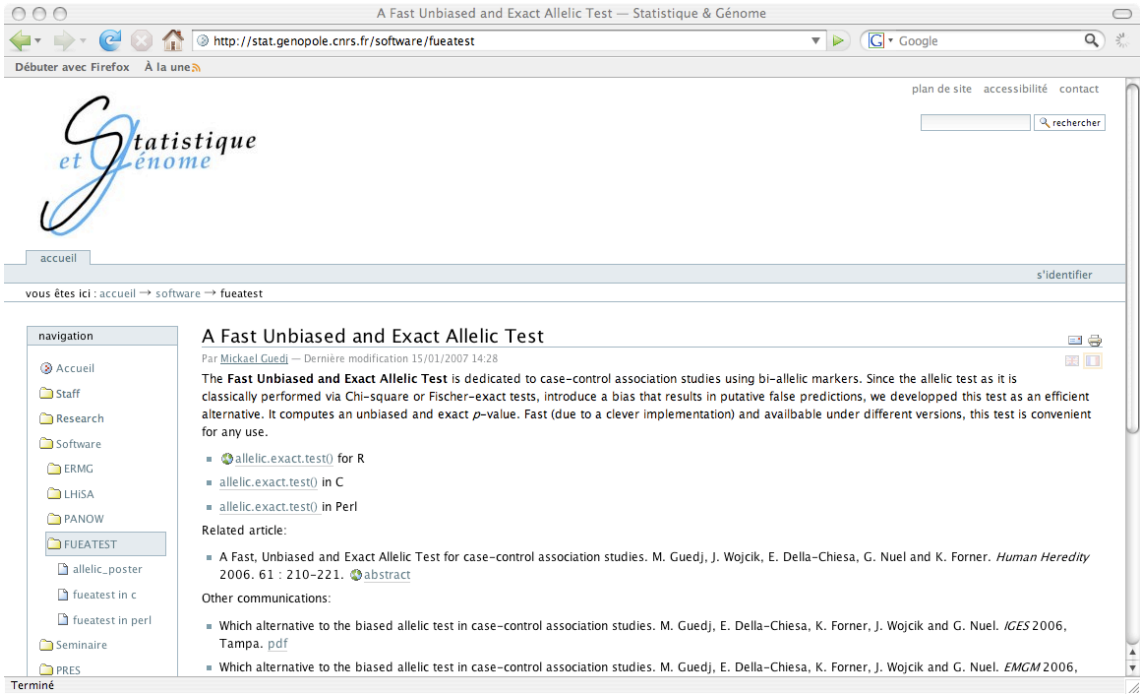


Fig. 2.8 – **Temps d'exécution :** estimé pour l'implémentation basique et optimisée du test non-biaisé sur 10,000 SNPs simulés en fonction de la taille de l'échantillon. Réalisé sur un processeur Itanium 2 (64 bits, 1.6 GHz).



`http://stat.genopole.cnrs.fr/fueatest`

- Comparaison qualitative : tout d'abord, la figure 2.9-A page 74 représente l'écart de significativité ($p\mathrm{v}_b - p\mathrm{v}_u$) généré par l'utilisation du test biaisé en fonction de la force de la déviation par rapport à l'équilibre d'Hardy-Weinberg, évaluée par la p-value du test d'Hardy-Weinberg appliqué sur les populations combinées de cas et de témoins ($p\mathrm{v}_{\mathrm{HW}}$). Comme attendu, les deux sont directement liés. Par ailleurs, la façon dont les allèles sont appariés dévie $p\mathrm{v}_b$ dans une direction prévisible : un déficit d'hétérozygotes augmente $p\mathrm{v}_b$ et par la même occasion le taux d'acceptation d'$H0$. En contrepartie, un excès d'hétérozygotes diminue $p\mathrm{v}_b$ et de fait augmente le taux d'erreur de type-I. La figure 2.9-B illustre le lien qu'il existe entre la fréquence allélique (p_A) et l'écart ($p\mathrm{v}_b - p\mathrm{v}_u$). Dans le cas d'un déficit d'hétérozygotes, l'on ne constate aucune relation évidente mais dans le cas d'un excès, l'écart maximum observé est clairement proportionnel à la quantité $p_A(1-p_A)$ et donc maximal pour $p_A = 0.5$ et minimal aux extrémités pour $p_A = 0$ et $p_A = 1$.

- Comparaison quantitative : une autre information intéressante est le pourcentage de prédictions erronées faites en utilisant le test biaisé (2.9-C). Ce pourcentage est simplement le rapport du nombre de faux (positifs et négatifs) générés par le test biaisé sur le nombre de positifs issus du test non biaisé. On constate qu'il est relativement important et augmente jusqu'à 9, 16, 30 et 40% lorsque l'on abaisse le niveau du test à 5, 1, 0.1 et 0.01% respectivement. Comme il est recommandé d'exclure d'une étude les SNPs qui ne répondent pas à l'équilibre d'Hardy-Weinberg chez les témoins (p. 32) nous avons également calculé ce pourcentage en rejetant les SNPs pour lesquels la p-value du test d'Hardy-Weinberg chez les témoins ne dépasse pas le seuil 5%, après correction de Benjamini-Hochberg pour le test-multiple (p. 30). Ce pourcentage est alors, certes moins important, mais atteint toujours un niveau de 28% pour un seuil fixé à 0.01% par exemple.

- Comparaison avec le test de tendance : enfin, puisque deux alternatives existent au test allélique biaisé, nous avons comparé les différences de puissance entre le test de tendance et le test allélique non-biaisé, en incluant ce dernier à l'étude de puissance présentée précédemment. Lorsque l'on fait varier la fréquence allélique (p_A), les puissances des deux tests ne sont pas significativement différentes (figure 2.10 p. 75). Elle sont d'ailleurs aussi identiques à celle du test allélique biaisé ce qui est cohérent avec le fait que lorsque la population combinée de cas et témoins est sous l'équilibre d'Hardy-Weinberg, les trois tests sont asymptotiquement équivalents. Quand on fait varier le coefficient de consanguinité ($\mathcal{F}$), le test de tendance et le test allélique non-biaisé affichent des puissances très comparables qui augmentent avec $\mathcal{F}$. Nous avons aussi représenté la puissance du test allélique biaisé bien qu'il n'y ait pas de sens de la comparer aux autres : en effet le taux d'erreur de type-I généré par ce test varie avec $\mathcal{F}$ ce qui se manifeste ici par une augmentation ($\mathcal{F} > 0$) et une diminution ($\mathcal{F} < 0$) artefactuelle de la puissance.

Par ailleurs malgré leur efficacité comparable, le test de tendance et le test non-biaisé présentent de petites mais significatives différences : si les deux sont identiques pour un MOI additif, le test de tendance est légèrement meilleur pour un MOI dominant et le test non-biaisé pour un MOI récessif. Ces différences sont accentuées pour un coefficient de consanguinité négatif ($\mathcal{F} < 0$) et augmentent avec le rapport cas-témoins (N_D/N_H).

Enfin en investigant l'effet de la taille de l'échantillon (N), on constate que celle-ci accentue clairement le biais introduit par le test allélique biaisé (figure 2.11 p. 76).

Discussion sur le test allélique :

Suite aux conclusions de Sasieni (1997), il est aujourd'hui bien connu que de comparer les proportions alléliques par l'intermédiaire d'un test du χ^2 ou un test exact de Fisher, est une stratégie biaisée du fait de l'appariement non-équiprobable des allèles. Le test conditionnel exact et optimisé que nous proposons constitue une solution au problème puisqu'il a l'avantage de ne faire aucune supposition sur la manière dont les allèles sont appariés. Sur la base de données *genome-wide*, nos résultats illustrent clairement que l'importance du biais dépend directement de la force de la déviation par rapport à l'équilibre d'Hardy-Weinberg. De plus la direction vers laquelle la p-value dévie est liée à la façon dont l'équilibre n'est pas respecté (excès ou déficit d'hétérozygotes). La taille de l'échantillon semble également avoir un impact, tout comme les proportions alléliques dans le cas d'un excès d'hétérozygotes. Enfin nous avons vu que l'impact du biais sur les prédictions augmente lorsque l'on diminue le niveau du test.

En pratique, ces conclusions ne jouent pas vraiment en faveur des études d'association : **(i)** le niveau de rejet d'$H0$ est généralement faible, traditionnellement fixé à 1% ou 5% et encore plus faible lorsque l'on prend en compte le test-multiple ; **(ii)** les études d'association font l'hypothèse que les allèles étiologiques ont une proportion conséquente dans la population en accord avec l'hypothèse *common disease - common variant*[17] et **(iii)** mettent aujourd'hui en jeu des tailles d'échantillons de plus en plus élevées. Par conséquent, si le réel impact lié à l'utilisation du test allélique biaisé n'est pas facile à estimer, il reste cependant une source d'erreur éventuelle pouvant diminuer la qualité des résultats d'une étude, et contre laquelle on sait aujourd'hui parfaitement se prémunir.

Comparé au test de tendance, le test non-biaisé présente des puissances tout à fait similaires sous différentes alternatives. Par conséquent le choix entre les deux tests se décide plutôt sur la base de considérations pratiques. On peut par exemple souligner le fait que, pour comparer des proportions alléliques, le test allélique reste sans doute pour le généticien plus intuitif que le test de tendance (Knapp 2003). Le test allélique non-biaisé constitue par ailleurs une extension naturelle du test biaisé, et nous le proposons sous différentes versions (R, C et Perl) afin de supporter sa diffusion. Enfin, avec l'accumulation des données en génétique, les analyses nécessitent aujourd'hui le traitement simultané d'un grand nombre de marqueurs (jusqu'à 500,000). Au delà des problèmes d'ordre statistique, le temps d'exécution reste une question importante. Les procédures de test exact ne sont pas réputées pour leur rapidité d'exécution et nous avons pensé important d'approfondir ce point : combinée à la puissance des machines actuelles, l'implémentation que nous proposons peut largement être adaptée à tout type de test d'association et rendre ainsi l'utilisation des tests exacts à grande échelle tout à fait envisageable.

[17] qui stipule que les polymorphismes conférant une susceptibilité à la maladie sont des variants communs dans la population. Cette hypothèse est largement discutée dans la communauté.

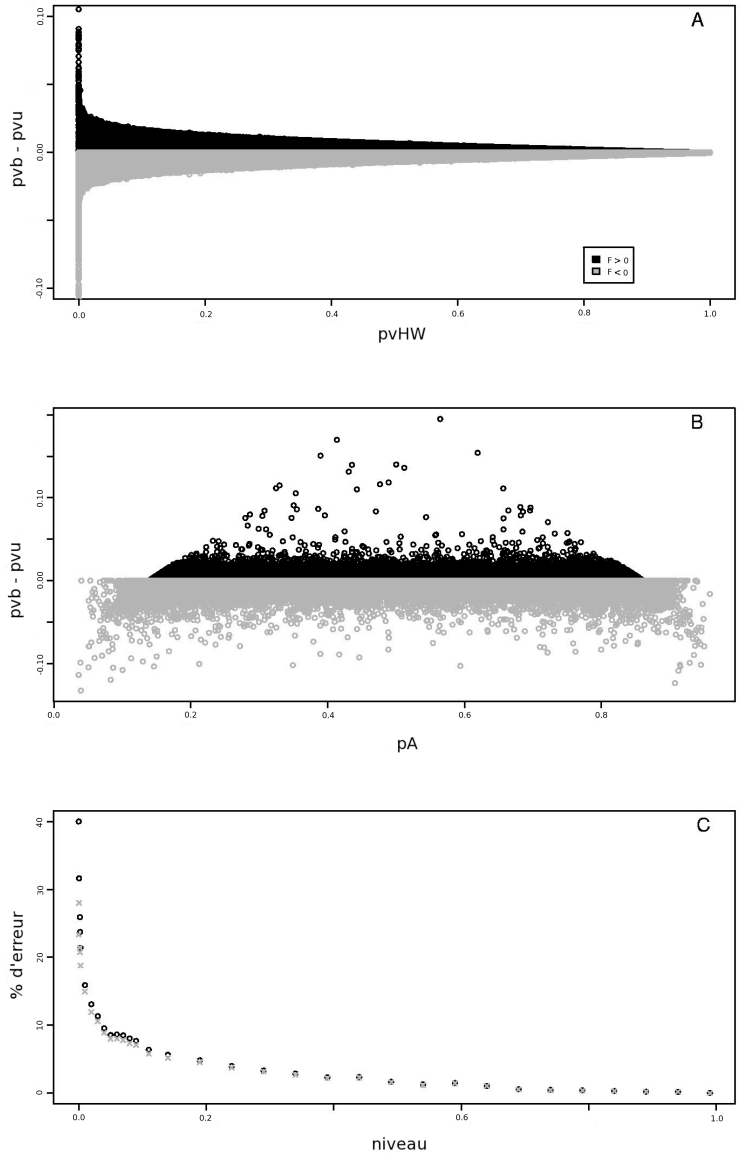


Fig. 2.9 – **Différence entre test allélique biaisé et non-biaisé sur les données AIM-Scan : A-** ce graphique représente l'erreur absolue faite en utilisant le test biaisé en fonction de la force de déviation par rapport à l'équilibre d'Hardy-Weinberg. **B-** ce graphique met en évidence le lien entre l'erreur faite en utilisant le test biaisé et la proportion allélique (p_A). **C-** Le pourcentage d'erreur est estimé par le rapport entre le nombre de faux-positifs et de faux-négatifs générés par l'utilisation du test biaisé sur le nombre de positifs obtenus avec le test non-biaisé. Il est représenté en fonction du niveau du test pour le jeu complet de SNPs (O) et le même jeu en excluant les SNPs qui ne respectent pas l'équilibre d'Hardy-Weinberg chez les témoins (X).

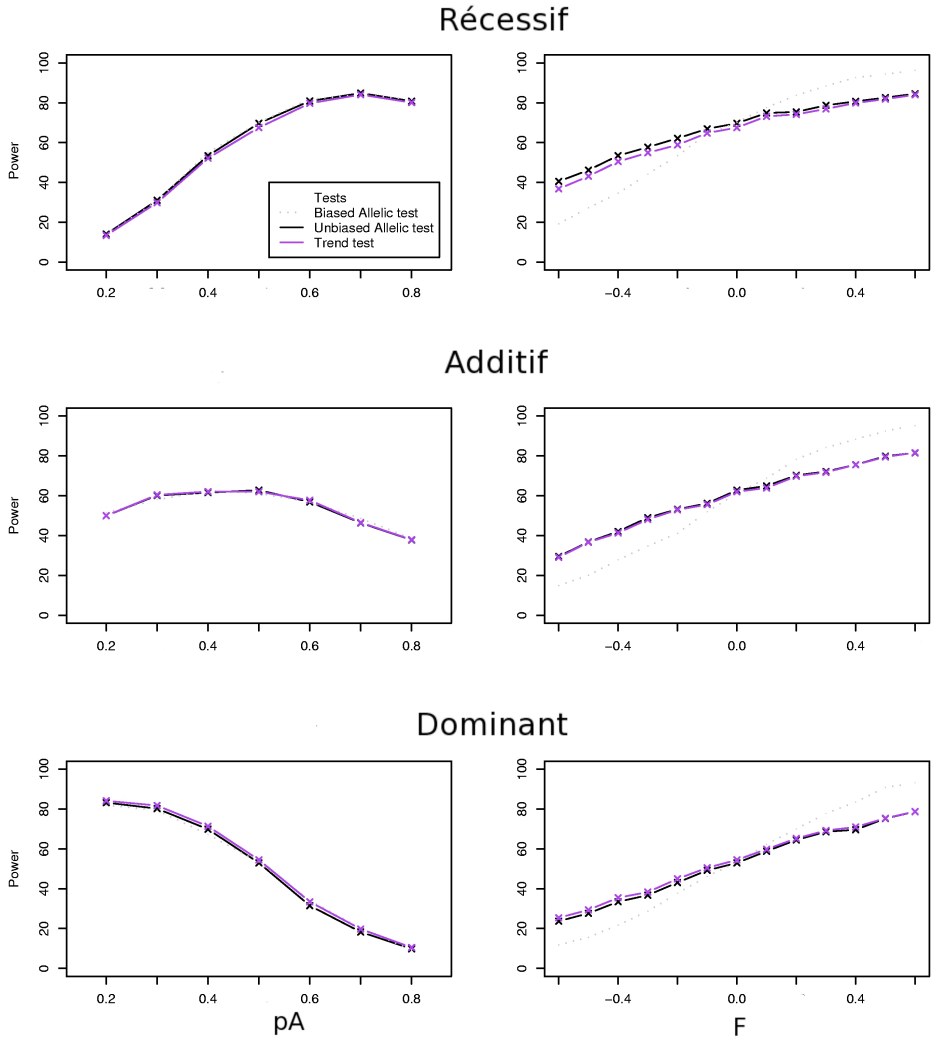


Fig. 2.10 – **Puissances du test allélique non-biaisé et du test de tendance :** en fonction de la proportion allélique p_A et du coefficient de consanguinité $\mathcal{F}$ au niveau 5% ($K_p = 0.05$, $RR_2 = 1.5$, $N_D = N_H = 500$).

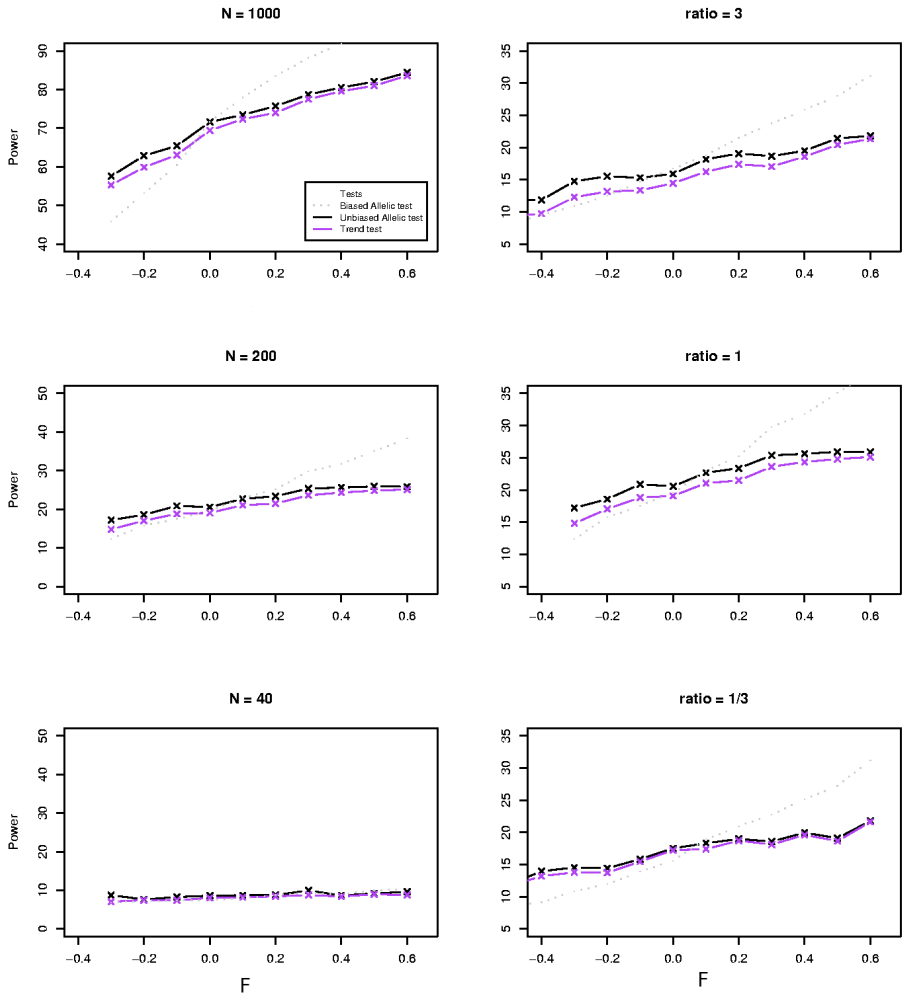


Fig. 2.11 – **Puissance du test allélique non-biaisé et du test de tendance :** en fonction du coefficient de consanguinité $\mathcal{F}$, de la taille d'échantillon N et du rapport de cas et de témoins N_D/N_H au niveau 5% ($K_p = 0.05$, $p_A = 0.5$, $RR_2 = 1.5$, MOI = R).

2.6 Cas particulier du test d'Hardy-Weinberg

État de l'art

Les études cas-témoins reposent sur le principe que l'allèle conférant une susceptibilité à la maladie devrait être plus représenté chez les cas. Cela implique par la même occasion des disproportions génotypiques par rapport à l'équilibre d'Hardy-Weinberg.

Constatant de telles disproportions dans leurs données, Feder et al (1996) suggèrent de tester l'équilibre chez les cas afin de contribuer à localiser les locus de susceptibilité. Par la suite, Nielsen et al (1999) proposent une généralisation du test à plus de deux allèles et en étudient certaines propriétés. La justification du test d'Hardy-Weinberg en tant que test d'association repose sur trois conditions d'application (voir annexe p. 192 pour leur justification) :

Condition 1 : la population générale est à l'équilibre.

Condition 2 : la pénétrance doit être incomplète.

Condition 3 : le mode de transmission sous-jacent ne doit pas être multiplicatif.

Par ailleurs le test présente deux propriétés intéressantes :

Propriété 1 : concernant trois des cinq modes de transmission pour lesquels on s'attend à trouver du déséquilibre (R, D, A), Nielsen et al (1999) ont montré que la puissance pour détecter le déséquilibre généré par la maladie décroît plus rapidement avec le LD que la puissance pour détecter de l'association avec un test classique d'association (*e.g.* génotypique, allélique ...).

Propriété 2 : une mesure d'association fondée sur le déséquilibre d'Hardy-Weinberg a l'avantage de nécessiter le recrutement de cas uniquement, ce qui peut s'avérer pratique lorsque des témoins ou des parents sont difficiles à obtenir.

- Fine-scale mapping : sur la base de la *Propriété* 1, les auteurs voient le test d'Hardy-Weinberg comme une façon de préciser la localisation d'un site étiologique. En particulier, Feder et al (1996) affichent le déséquilibre chez les cas par chacun de leur marqueur. La courbe obtenue présente approximativement le même maximum que la courbe obtenue avec leur mesure d'association (p_{excess} équivalent à une différence entre proportions alléliques $p_{D_A} - p_{D_a}$), avec cependant une forme plus pointue (figure 2.12 p. 78). Les auteurs en conclurent la confirmation de leur résultat préliminaire et la précision de la région de susceptibilité. Nielsen et al (1999) confirmèrent cette propriété analytiquement.

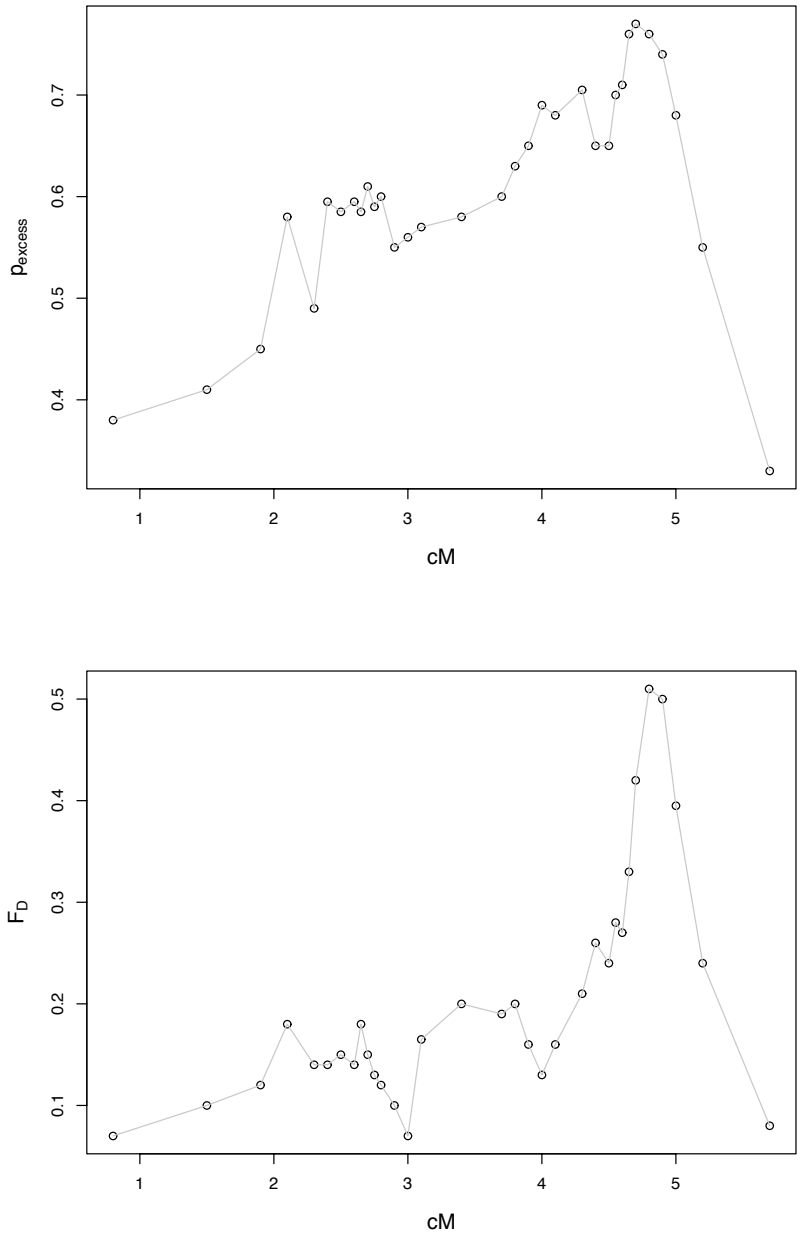


Fig. 2.12 – **Étude de Feder et al :** les deux graphiques illustrent les résultats obtenus par les auteurs en affichant pour le même jeu de marqueurs, une mesure de différence entre les proportions alléliques observées chez les cas et chez les témoins ($p_{\text{excess}} = p_{D_A} - p_{D_a}$) et une mesure du déséquilibre d'Hardy-Weiberg observé chez les cas ($\mathcal{F}_D$).

- *Genome-wide screening tool* : plus tard, Lee (2003) laisse cette idée de côté et propose d'utiliser le test d'Hardy-Weinberg comme outils d'investigation *genome-wide* de banques d'individus affectés (*Propriété* 2). Il montre en particulier que cette approche est plus puissante dans le cas d'un mode de transmission récessif ou dominant et à nombre d'individus génotypés égal, qu'une approche familiale utilisant le *Transmission Disequilibrium Test* (Spielman et al 1993).

- Limitations : le test d'Hardy-Weinberg tel qu'il est proposé par Feder et al (1996) repose sur une condition d'application très contraignante : la population générale doit être à l'équilibre (*Condition* 1). Comme déjà abordé dans ce manuscrit, faire cette hypothèse n'est pas nécessairement judicieux, l'équilibre pouvant être mis à mal pour différentes raisons (p. 12 et 32). L'utilisation du test lorsque cette condition n'est pas respectée peut donc être à l'origine de faux-positifs. Par ailleurs, la *Condition* 3 limite le champ de son application à des situations qui ne peuvent être prévues à l'avance. Par conséquent, l'utilisation du test d'Hardy-Weinberg en tant que test d'association dans les études *genome-wide* est sujet à controverse.

Erreur de type-I et puissance

L'inflation du taux d'erreur de type-I et l'insuffisance de puissance du test d'Hardy-Weinberg (X_{HW}) n'ont pas été étudiés de façon exhaustive. Il n'est pas clair comment ces quantités se comportent en fonction de paramètres du modèle génétique tels que la proportion alléliques (p_A), du coefficient de consanguinité ($\mathcal{F}$) et du déséquilibre de liaison entre le marqueur considéré et le locus étiologique (r^2). Nous avons inclu ce test à notre étude de puissance de façon à confirmer, discuter et compléter les résultats observés sur des études indépendantes. En plus de comparer le test d'Hardy-Weinberg aux tests de tendance (X_T) et génotypique (X_G), nous introduisons une alternative qui doit permettre de se prémunir des faux-positifs lorsque la population générale n'est pas à l'équilibre. Ce test a déjà été décrit par Song et al (2005) en tant que *Hardy-Weinberg Disequilibrium Trend Test*. Il compare le déséquilibre d'Hardy-Weinberg observé chez les cas et les témoins et postule qu'une différence significative doit être due à la maladie :

$$H0 : \{\mathcal{F}_D = \mathcal{F}_H\}.$$

La statistique correspondante mesure donc une distance entre le déséquilibre observé chez les cas et chez les témoins :

$$\Delta_{\mathcal{F}} = F_D - F_H \underset{H0}{\sim} \mathcal{N}(0, 2).$$

La figure 2.13 page 82 présente le taux d'erreur de type-I généré par chacun des tests pour un niveau de 5%. Pour $\mathcal{F} = 0$, cette quantité demeure aux alentours de 5% pour tous les tests. Lorsque la population générale s'écarte de l'équilibre, le taux d'erreur de type-I

augmente rapidement à 1 pour X_{HW} alors qu'il est contrôlé à la valeur attendue pour les trois autres tests (X_T, X_G et $\Delta_{\mathcal{F}}$).

Les figures 2.14 et 2.15 p. 83 et 84 indiquent la puissance de chaque test en fonction des proportions alléliques et du déséquilibre de liaison. On voit que lorsque le coefficient de consanguinité est nul dans la population générale et malgré son contrôle du taux d'erreur de type-I, $\Delta_{\mathcal{F}}$ est généralement le moins puissant. Ce résultat n'est pas surprenant, ce test possédant les défauts de ses qualités : lorsqu'on ne connaît pas le coefficient de consanguinité, il permet de l'estimer et de maintenir ainsi le taux d'erreur de type-I au niveau souhaité ; lorsque l'on sait que ce coefficient est nul, chercher à l'estimer entraîne obligatoirement une diminution de puissance. Pour de faibles valeurs de p_A (< 0.4) et un mode de transmission récessif, X_{HW} est le plus puissant. De façon symétrique il l'est aussi pour de fortes valeurs de p_A (> 0.6) et un mode de transmission récessif. Par ailleurs on peut noter que ce test est également le meilleur à percevoir un mode de transmission sur-dominant pour des variants communs dans la population ($0.2 < p_A < 0.8$).

En ce qui concerne le mode de transmission additif/multiplicatif, les puissances des deux tests fondés sur l'équilibre d'Hardy-Weinberg (X_{HW} et $\Delta_{\mathcal{F}}$) restent au niveau du taux d'erreur de type-I fixé (ici 5%). Cela s'explique par le fait que ces tests ne sont pas capables de capter des situations générées à partir d'un mode de transmission multiplicatif (*Condition* 3) ou du moins proche. Enfin la décroissance plus prononcée de la puissance de X_{HW} avec r^2 (*Propriété* 2) s'observe également pour un mode de transmission sur-dominant (mis en avant sur la figure 2.15) ce qui n'avait pas encore été constaté ; en revanche, le test fondé sur $\Delta_{\mathcal{F}}$ ne possède pas cette propriété contrairement à ce que certains auteurs semblent affirmer sans apporter cependant aucune justification (Song et Elston 2006).

Équilibre d'Hardy-Weinberg et méta-statistiques

L'équilibre d'Hardy-Weinberg n'a pas échappé à la tentation d'être combiné à d'autres statistiques avec l'intention d'y gagner en puissance. Dans leur *Set Association*, Hoh et al (2001) pondèrent la statistique du test allélique X_A par la statistique du test d'Hardy-Weinberg calculé chez les contrôles mais cette approche s'est révélée moins puissante que le test allélique réalisé seul (Hao et al 2004). Song et Elston (2006) proposent une combinaison linéaire de X_T avec $\Delta_{\mathcal{F}}$: *the Weighted Average statistic*. Dans leurs simulations, ce test est globalement meilleur que le test de tendance seul.

Ces résultats corroborent ceux de notre étude de puissance où le fait d'ajouter X_{HW} aux statistiques X_T (figure 2.16 p. 85) et X_{T+G} (figure 2.17 p. 86) pour donner $X_{T+\mathrm{HW}}$ et $X_{T+G+\mathrm{HW}}$ respectivement permet une augmentation de puissance dans la plupart des situations. Les bonnes performances de l'ajout de X_{HW} sont en quelque sorte surprenantes compte tenu des précédentes conclusions sur les méta-statistiques (p. 61) qui tendent à montrer qu'elles n'ont pas pour effet d'y gagner en puissance.

Ces observations peuvent s'expliquer par la nature particulière du test d'Hardy-Weinberg. Contrairement aux tests génotypique, allélique ou de tendance, il n'utilise pas à proprement parler l'information d'association entre le marqueur et la maladie, mais se sert plutôt du déséquilibre gamétique observé chez les cas comme substitut de cette association. Il utilise donc une information différente pour détecter le même phénomène et l'on peut se demander si l'augmentation de puissance observée n'est pas due à une indépendance des tests fondés sur Hardy-Weinberg avec les autres tests d'association. Des estimations empiriques des covariances entre le test d'Hardy-Weinberg et les tests génotypique, allélique et de tendance ont clairement rejeté l'hypothèse d'indépendance avec le test génotypique, bien que la covariance soit faible. En revanche la covariance estimée nulle entre le test d'Hardy-Weinberg et les tests allélique et de tendance ne nous permet pas de rejeter cette hypothèse. Cela ne constitue pas en revanche une preuve d'indépendance et d'un point de vue technique la preuve formelle reste à proposer. Néanmoins, pour nous en convaincre, nous avons évalué empiriquement la distribution sous $H0$ de la statistique $X_{T+\mathrm{HW}}$. Sous l'hypothèse d'indépendance entre X_T et X_{HW}, $X_{T+\mathrm{HW}}$ suit asymptotiquement une distribution du χ^2 à 2 degrés de liberté ce que semble nous révéler la figure 2.13-B page 82 où les deux distributions se calquent parfaitement l'une sur l'autre.

On constate par ailleurs que l'augmentation de puissance ne concerne pas le cas additif/multiplicatif (figure 2.16 p. 85 et figure figure 2.17 p. 86). Le test d'Hardy-Weinberg étant en effet inefficace dans ce type de situation, on aurait pu s'attendre à ce que les puissance de $X_{T+\mathrm{HW}}$ et $X_{T+G+\mathrm{HW}}$ soit au niveau de celles X_T et X_{T+G} respectivement. Mais le fait d'ajouter X_{HW} introduit un bruit qui va avoir pour effet de dégrader le signal plutôt que de l'améliorer. Notons néanmoins que cette diminution de puissance reste minime comparée à l'augmentation globale dans les autres modes de transmission.

Discussion sur le test d'Hardy-Weinberg

A première vue, le test d'Hardy-Weinberg appliqué sur les cas possède des propriétés intéressantes. En particulier il permet de détecter de l'association en présence de cas uniquement (*Propriété* 2) ce qui peut être très avantageux d'un point du vue du *design* de l'étude. Par ailleurs, sous certains modèles génétiques, il permet de préciser la localisation du locus étiologique (*Propriété* 1). Néanmoins, les restrictions liées à son utilisation ne nous permette pas de le recommander pour les études d'association *genome-wide*. En particulier, il fait l'hypothèse forte que la population générale est à l'équilibre, ce qui peut entraîner une augmentation inquiétante du taux de faux-positifs lorsque cette hypothèse n'est pas réalisée. Par ailleurs ce test est performant dans des modes de transmission limites tels que récessif, dominant et sur-dominant (R, D et S). Mais il est totalement inefficace face à des situations intermédiaires telles que les modes de transmission additifs et multiplicatifs (A et M). De telles situations ont pourtant toutes les chances d'être la règle plutôt que des cas particuliers : les études d'association *genome-wide* reposent essentiellement sur l'association indirecte entre les marqueurs et les locus de susceptibilité. Du fait du déséquilibre de liaison, ces marqueurs présentent vraisemblablement une relation

au risque d'être malade intermédiaire entre les modes de transmission limites (dominant, récessif) même si les locus pour lesquels ils servent de substitut présentent eux-mêmes un de ces modes de transmission (Weinberg and Morris 2003).

L'alternative que nous proposons et qui a été précédemment décrite par Song et al (2005), repose sur la statistique $\Delta_{\mathcal{F}}$ et permet de contrôler efficacement le taux de faux-positifs au niveau α fixé, quel que soit le coefficient de consanguinité dans la population générale. Néanmoins ce test s'avère bien moins puissant que le test d'Hardy-Weinberg et demeure inefficace pour les modes de transmission additifs et multiplicatifs. Par ailleurs, il ne possède aucune des deux propriétés intéressantes du test d'Hardy-Weinberg. Par conséquent il ne constitue *a priori* pas une alternative réellement satisfaisante comparée aux tests d'association plus classiques tels que le test génotypique ou le test de tendance.

En revanche, combinée à d'autres statistiques d'association, l'information apportée par le déséquilibre d'Hardy-Weiberg observé chez les cas devient pertinente. Certains auteurs observent une augmentation sensible de puissance (Song et Elston 2006) ce que nous confirmons par notre étude de puissance. Néanmoins, et comme nous l'avons déjà précisé, l'utilisation de tels tests requiert une certaine attention quant à l'estimation de la p-value du fait de la dépendance entre les statistiques. Une indépendance entre X_T et X_{HW} permettrait d'ailleurs d'approcher directement la distribution de la statistique $X_{T+\mathrm{HW}}$ par un χ^2 à 2 degrés de liberté. Si c'est en pratique ce que nous observons, cette indépendance n'a pas été formellement démontrée, ce qui constitue un point technique à compléter par la suite.

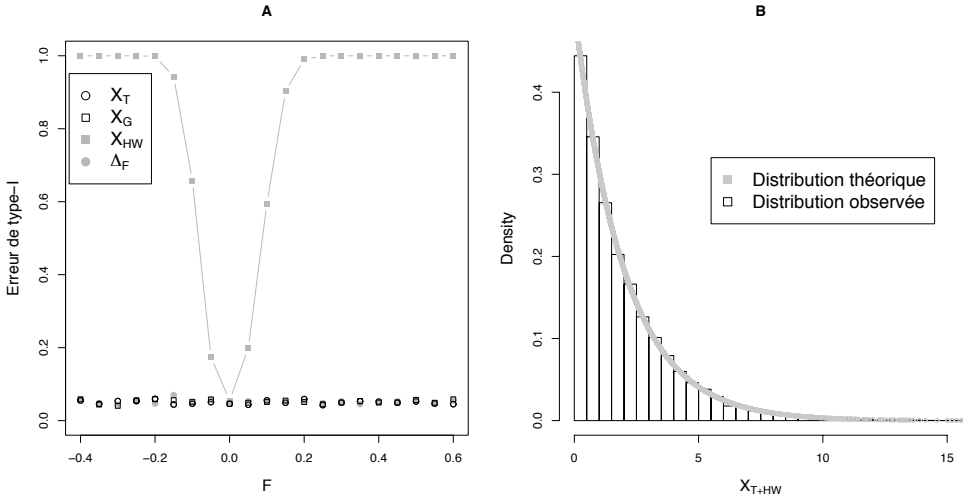


Fig. 2.13 – **A - Erreur de type-I de 4 tests d'association :** X_T, X_G, X_{HW} et $\Delta_{\mathcal{F}}$ en fonction du coefficient de consanguinité $\mathcal{F}$ au niveau 5% ($K_p = 0.05$, $p_A = 0.4$, $\mathcal{F} = 0$, $RR_2 = 1$, $RR_1 = 1$ et $N_D = N_H = 500$). **B - Distribution observée de** $X_{T+\mathrm{HW}}$ **:** comparaison avec la distribution théorique sous l'hypothèse d'indépendance entre X_T et X_{HW} : $\chi^2(2)$. Test d'adéquation des deux distributions de Kolmogorov-Smirnov : $p\mathrm{v} \simeq 0.8$.

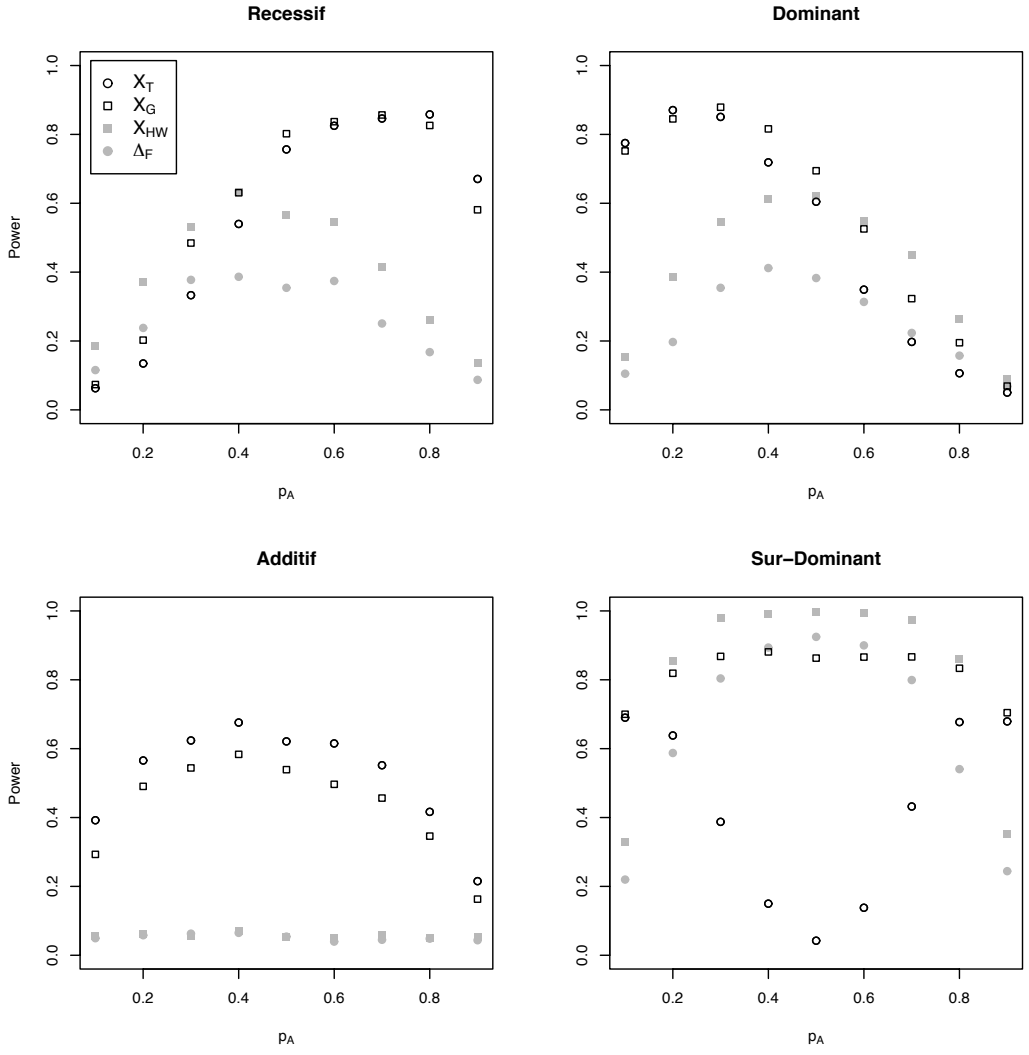


FIG. 2.14 – **Puissance de 4 tests d'association :** X_T, X_G, X_{HW} et $\Delta_{\mathcal{F}}$ en fonction des proportions alléliques p_A au niveau 5% ($K_p = 0.05$, $\mathcal{F} = 0$, $RR_2 = 1.5$ pour les modes de tranmission récessif, additif et dominant ou $RR_1 = 1.5$ pour le mode de transmission sur-dominant, $N_D = N_H = 500$).

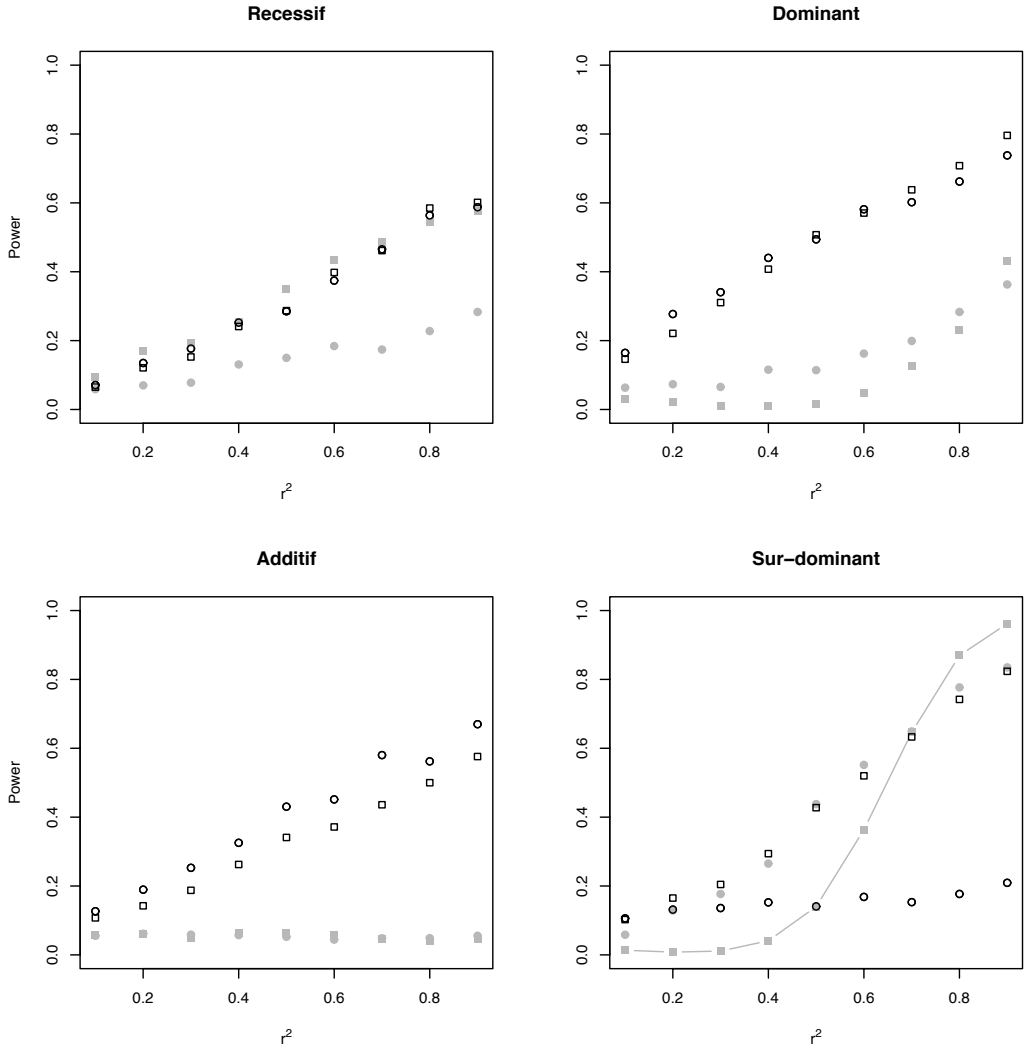


Fig. 2.15 – **Puissance de 4 tests d'association :** X_T, X_G, X_{HW} et $\Delta_{\mathcal{F}}$ en fonction du déséquilibre de liaison r^2 au niveau 5% ($K_p = 0.05$, $p_A = 0.4$, $\mathcal{F} = 0$, $RR_2 = 1.5$ pour les modes de transmissions récessif, additif et dominant ou $RR_1 = 1.5$ pour le mode de transmission sur-dominant, $N_D = N_H = 500$). La décroissance prononcée de la puissance associée à X_{HW} dans le modèle sur-dominant est mise en avant par un trait.

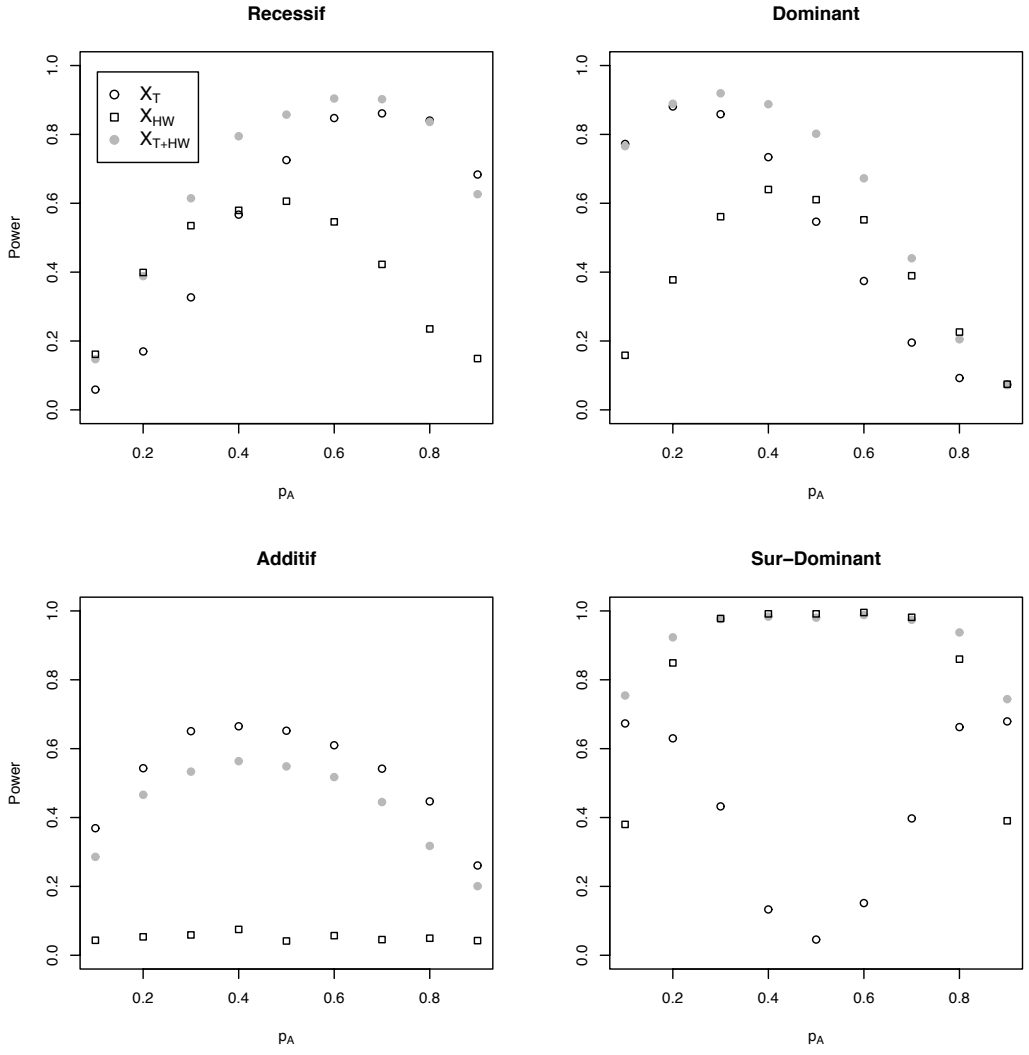


FIG. 2.16 – **Puissance des méta-statistiques impliquant Hardy-Weinberg :** X_T, X_{HW} et X_{T+HW} en fonction des proportions alléliques p_A, au niveau 5% ($K_p = 0.05$, $\mathcal{F} = 0$, $RR_2 = 1.5$ pour les modes de transmission récessif, additif et dominant ou $RR_1 = 1.5$ pour le mode de transmission sur-dominant, $N_D = N_H = 500$).

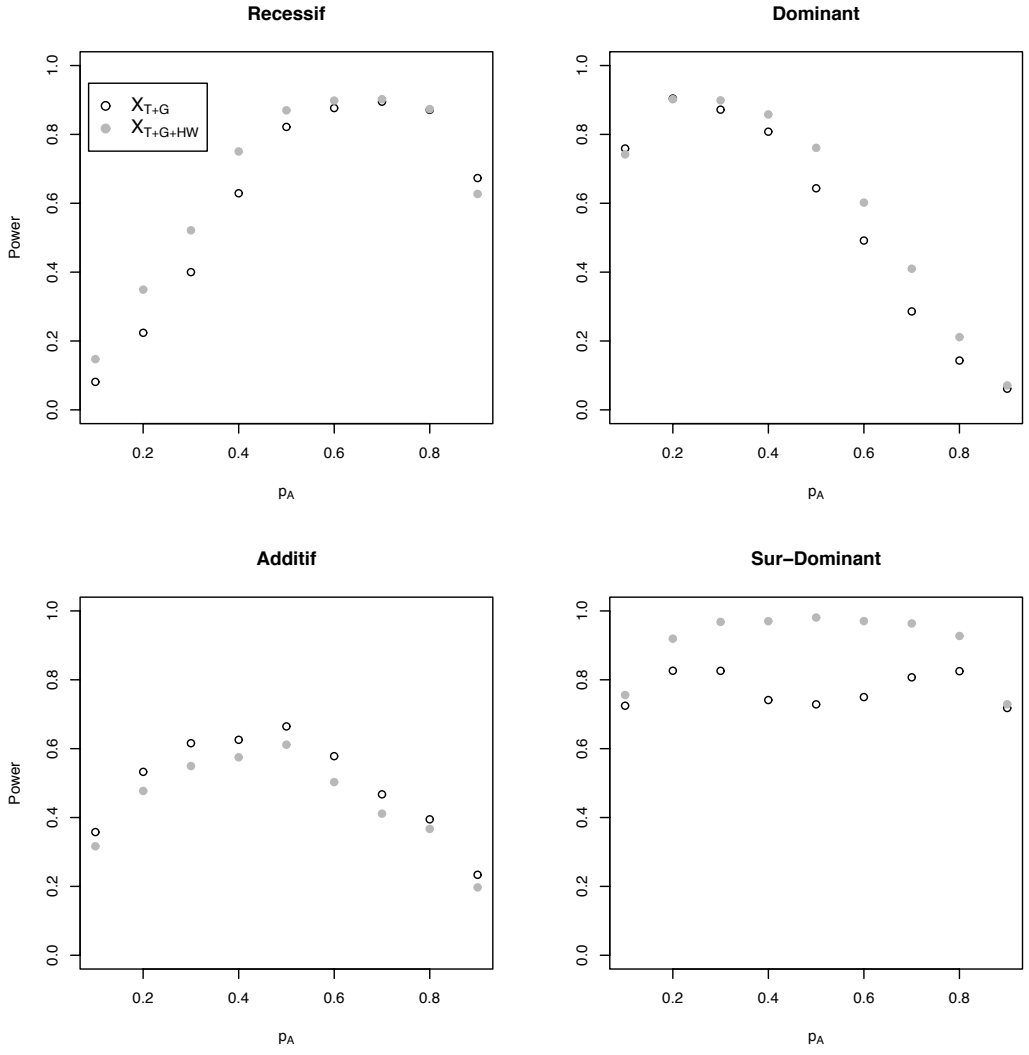


Fig. 2.17 – **Puissance des méta-statistiques impliquant Hardy-Weinberg :** X_{A+G} et $X_{A+G+\mathrm{HW}}$ en fonction des proportions alléliques p_A, au niveau 5% ($K_p = 0.05$, $\mathcal{F} = 0$, $RR_2 = 1.5$ pour les modes de transmission récessif, additif et dominant ou $RR_1 = 1.5$ pour le mode de tranmission sur-dominant, $N_D = N_H = 500$).

2.7 FDR Local

Comme nous l'avons déjà évoqué, la dimension des jeux de données générés aujourd'hui implique de réaliser un grand nombre de tests d'hypothèse à un niveau α, ce qui conduit au problème du test-multiple lié à une accumulation de faux-positifs (introduction p. 30). Comme alternative à un contrôle du FWER[18] souvent considéré inadapté au problème, le contrôle du FDR a été introduit par Benjamini et Hochberg (1995). Néanmoins, un inconvénient du FDR est qu'il est attaché à une région de rejet (traditionnellement $[0;\alpha]$) et par conséquent à l'ensemble des marqueurs dont la p-value se trouvent dans cette région. Il ne prend donc pas en compte la proximité d'une p-value vers les limites 0 ou α, qui fait pourtant varier la probabilité pour le marqueur correspondant d'être sous $H0$ ou $H1$. Afin de répondre à ce problème, Efron (2001) a récemment introduit la notion de FDR Local noté fdr, et qui réfère à la probabilité spécifique pour chaque marqueur d'être sous $H0$, c'est à dire de ne pas être associé à la maladie.

Après avoir posé la définition du FDR Local, nous proposons à travers un modèle de mélange gaussien, un cadre naturel, simple et didactique pour aborder son estimation. Le modèle en question a précédemment été employé dans l'étude de données d'expression de gènes (McLachlan et al 2006). A la suite d'une application sur des simulations et sur des données réelles *genome-wide*, nous discutons les avantages et inconvénients de cette approche.

Définition

Le FDR Local provient d'un formalisme bayésien du problème du test-multiple. Soit ($p\mathrm{v}_1$, ..., $p\mathrm{v}_n$) l'ensemble des p-values associées aux n marqueurs. Ces p-values tombent dans deux classes selon que p_i ait été générée sous l'hypothèse $H0$ ou $H1$. Par conséquent, la distribution de l'ensemble des p-values (dont f est la fonction densité) résulte d'un mélange entre les p-values générées sous $H0$ et celles générées sous $H1$:

$$f = \pi_0 f_0 + \pi_1 f_1,$$

avec f_0 et f_1 les fonctions densité des p-values et π_0 et π_1 les probabilités *a priori* d'être sous $H0$ et sous $H1$ respectivement. A partir du théorème de Bayes, Efron fournit la probabilité *a priori* d'être sous $H0$ sachant la valeur $p\mathrm{v}_i$ de la p-value :

$$\mathrm{fdr}(p\mathrm{v}_i) \equiv \frac{\pi_0 f_0(p\mathrm{v}_i)}{f(p\mathrm{v}_i)}.$$

Si l'on note F_0, F_1 et F les fonctions de répartition (CDF) associées aux densités f_0, f_1 et f respectivement, l'on obtient une expression du FDR pour un niveau α fixé suivant le même formalisme :

$$\mathrm{FDR}(\alpha) \equiv \frac{\pi_0 F_0(\alpha)}{F(\alpha)} = \frac{\pi_0 \int_0^\alpha f_0(x)dx}{\int_0^\alpha f(x)dx}.$$

[18] type correction de Bonferroni

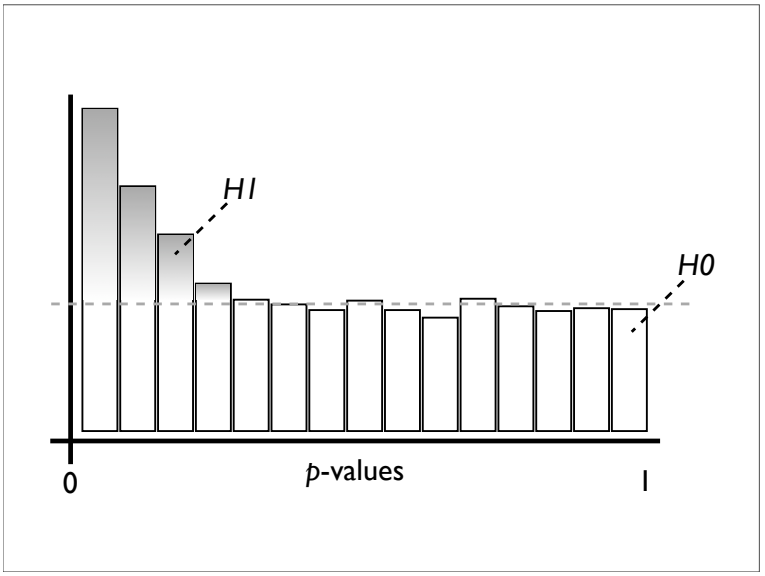


Fig. 2.18 – **Mélange de distribution des p-values.**

Estimation du FDR Local par algorithme EM

Suivant le formalisme précédemment, l'estimation du FDR Local se réduit à une estimation de densités dans un modèle de mélange. Une approche très simple et naturelle de procéder est de passer par un algorithme EM dont une introduction est proposée en annexe p. 193.

- Généralisation du problème : en toute généralité, soit $\mathbf{X} = (X_1, \cdots, X_n)$ un vecteur aléatoire des observations pour lequel les X_i sont des variables aléatoires indépendantes, de réalisation x_i et de fonction densité $f(x_i)$ résultant d'un mélange de C classes telles que :

$$f_{\pi,\theta}(x_i) = \sum_{c=1}^{C} \pi_c f_{\theta_c}(x_i).$$

On peut noter que $\pi = (\pi_1, \cdots, \pi_C)$ avec $\sum \pi_c = 1$ et $\theta = (\theta_1, \cdots, \theta_C)$ sont les paramètres du mélange. De façon pratique, π_c est tout simplement la probabilité *a priori* pour toute observation d'appartenir à la classe c. Formulé ainsi, le modèle présente un problème de données incomplètes puisque le classe à laquelle appartient chaque observation est inconnue. Soit $Z = (Z_1, \cdots, Z_n)$ un vecteur aléatoire de données incomplètes pour lequel les $Z_i = (Z_{i1}, \cdots, Z_{iC})$ sont des vecteurs aléatoires indépendants de taille C et de réalisation z_i avec z_{ic} prenant la valeur 1 si x_i appartient à la classe c et 0 sinon. La fonction densité des données complètes $f^{\mathbf{c}}$ de paramètres π et θ est alors exprimée par :

$$f^{\mathbf{c}}_{\pi,\theta}(x_i, z_i) = \sum_{c=1}^{C} z_{ic} \times \pi_c f_{\theta_c}(x_i).$$

Dans ce contexte, une fois que les estimations des paramètres au maximum de vraisemblance ($\widehat{\pi}$ et $\widehat{\theta}$) sont connues, il est possible d'estimer la probabilité *a posteriori* (τ_{ic}) pour chaque observation (x_i) d'appartenir à la classe c :

$$\widehat{\tau}_{ic} = \frac{\widehat{\pi}_c f_{\widehat{\theta}_c}(x_i)}{\sum_{k=1}^{C} \widehat{\pi}_k f_{\widehat{\theta}_k}(x_i)}.$$

Comme indiqué en annexe p. 193, à l'itération h de l'algorithme EM, l'étape E détermine l'espérance conditionnelle $\mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\{\log \mathbb{P}(\mathbf{X},\mathbf{z}|\pi,\theta)\}$:

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\{\log \mathbb{P}(\mathbf{X},\mathbf{z}|\pi,\theta)\} &= \mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\left\{\sum_{i=1}^{n} \log f^{\mathbf{c}}_{\pi,\theta}(x_i,z_i)\right\} \\
&= \mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\left\{\sum_{i=1}^{n} \log\left(\sum_{c=1}^{C} z_{ic}\pi_c f_{\theta_c}(x_i)\right)\right\} \\
&= \mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\left\{\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}\log\left(\pi_c f_{\theta_c}(x_i)\right)\right\} \\
&= \sum_{i=1}^{n}\sum_{c=1}^{C} \mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\left\{z_{ic}\log\left(\pi_c f_{\theta_c}(x_i)\right)\right\} \\
&= \sum_{i=1}^{n}\sum_{c=1}^{C} \mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\left\{z_{ic}\right\}\log\left(\pi_c f_{\theta_c}(x_i)\right)
\end{aligned}$$

Avec :

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\left\{z_{ic}\right\} &= \mathbb{P}(Z_{ic}=1|X_i=x_i,\pi^{(h)},\theta^{(h)}) \\
&= \tau_{ic}^{(h+1)} \\
&= \frac{\pi_c^{(h)} f_{\theta_c^{(h)}}}{\sum_{k=1}^{C} \pi_k^{(h)} f_{\theta_k^{(h)}}(x_i)}
\end{aligned}$$

D'où :

$$\mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\{\log \mathbb{P}(\mathbf{X},\mathbf{z}|\pi,\theta)\} = \sum_{i=1}^{n}\sum_{c=1}^{C} \tau_{ic}^{(h)} \log\left(\pi_c f_{\theta_c}(x_i)\right).$$

L'étape M maximise cette expression par rapport aux paramètres π et θ afin de donner une mise à jour de leurs estimations $\pi^{(h+1)}$ et $\theta^{(h+1)}$. Dans un premier temps, on peut estimer $\pi^{(h+1)}$ à partir de la contribution de chaque observation à chaque classe *via* les probabilités *a posteriori* $\tau_{ic}^{(h+1)}$:

$$\widehat{\pi}_c^{(h+1)} = \frac{\sum_{i=1}^{n} \tau_{ic}^{(h+1)}}{n}.$$

Enfin l'estimation de θ à l'itération $(h+1)$ est obtenue en résolvant l'équation :

$$\frac{\partial \mathbb{E}_{\mathbf{z}|\mathbf{X},\pi^{(h)},\theta^{(h)}}\{\log \mathbb{P}(\mathbf{X},\mathbf{z}|\pi,\theta)\}}{\partial\theta} = 0.$$

- Application à l'estimation du FDR Local : ici, le vecteur des observations est l'ensemble des p-values $(p\text{v}_1, \cdots, p\text{v}_n)$. Le nombre de classes C est égal à 2 pour les p-values générées sous $H0$ et celles générées sous $H1$. Les paramètres du modèle sont donc $\pi = (\pi_0, \pi_1)$ avec $\pi_1 = 1 - \pi_0$ et $\theta = (\theta_0, \theta_1)$. Sous cette formulation, les fonctions densités du modèle de mélange f_{θ_0} et f_{θ_1} ne sont pas évidentes à exprimer ; une façon plus confortable de traiter le problème consiste à appliquer une transformation probit sur les p-values de sorte que les fonctions densités sous $H0$ et $H1$ soient celles de distributions normales $\mathcal{N}(\mu_0, \sigma_0)$ et $\mathcal{N}(\mu_1, \sigma_1)$ de paramètres $\theta_0 = (\mu_0, \sigma_0)$ et $\theta_1 = (\mu_1, \sigma_1)$ respectivement :

$$x_i = \text{probit}(p\text{v}_i) = \Phi^{-1}(p\text{v}_i),$$

$$f_{\theta_j}(x_i) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{\frac{-(x_i - \widehat{\mu}_j)}{2(\sigma_j)^2}},$$

avec Φ la CDF d'une distribution normale centrée réduite. L'application de l'algorithme EM déroulé précédemment permet d'obtenir les estimations au maximum de vraisemblance des τ_{i0} et τ_{i1}. A l'itération (h) on obtient une estimation de $\tau_{i0}^{(h+1)}$ et $\tau_{i1}^{(h+1)}$ par :

$$\widehat{\tau}_{i0}^{(h+1)} = \frac{\widehat{\pi}_0^{(h)} f_{\theta_0}(x_i)}{\widehat{\pi}_0^{(h)} f_{\theta_0}(x_i) + \widehat{\pi}_1^{(h)} f_{\theta_1}(x_i)},$$

$$\widehat{\tau}_{i1}^{(h+1)} = \frac{\widehat{\pi}_1^{(h)} f_{\theta_0}(x_i)}{\widehat{\pi}_0^{(h)} f_{\theta_0}(x_i) + \widehat{\pi}_1^{(h)} f_{\theta_1}(x_i)},$$

Ensuite l'on obtient une estimation des paramètres π_0, π_1, θ_0 et θ_1 par :

$$\begin{aligned}
\pi_0^{(h+1)} &= \frac{\sum_{i=1}^n \tau_{i0}^{(h+1)}}{n}, \\
\pi_1^{(h+1)} &= 1 - \pi_0^{(h+1)}, \\
\mu_0^{(h+1)} &= \frac{\sum_{i=1}^n x_i \tau_{i0}^{(h+1)}}{\sum_{i=1}^n \tau_{i0}^{(h+1)}}, \\
\mu_1^{(h+1)} &= \frac{\sum_{i=1}^n x_i \tau_{i1}^{(h+1)}}{\sum_{i=1}^n \tau_{i1}^{(h+1)}}, \\
\sigma_0^{(h+1)} &= \frac{\sum_{i=1}^n x_i^2 \tau_{i0}^{(h+1)}}{\sum_{i=1}^n \tau_{i0}^{(h+1)}} - (\mu_0^{(h+1)})^2, \\
\sigma_1^{(h+1)} &= \frac{\sum_{i=1}^n x_i^2 \tau_{i1}^{(h+1)}}{\sum_{i=1}^n \tau_{i1}^{(h+1)}} - (\mu_1^{(h+1)})^2.
\end{aligned}$$

L'on obtient ainsi directement une estimation du FDR Local en remarquant que $\text{fdr}(p\text{v}_i)$ et τ_{0i} estiment tous deux la probabilité *a posteriori* d'être sous $H0$ sachant la valeur de la $i$ème observation :

$$\text{fdr}(p\text{v}_i) = \tau_{0i}.$$

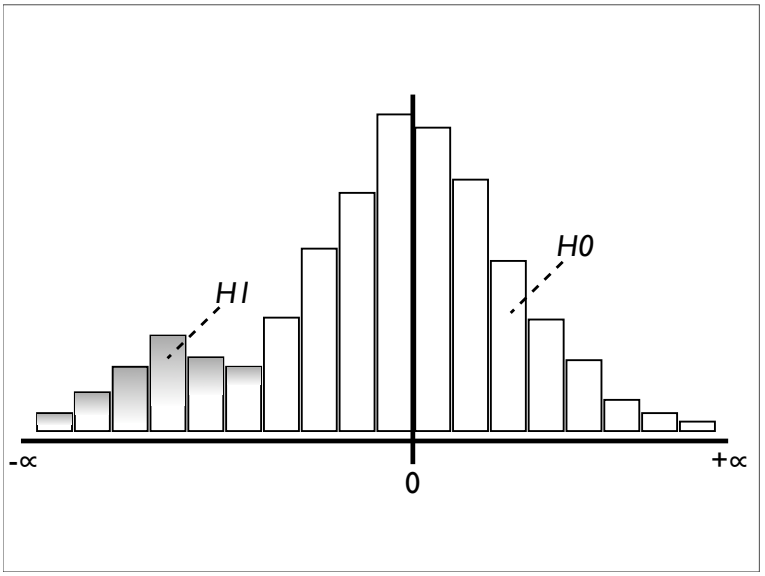


FIG. 2.19 – **Mélange de distribution des p-values après transformation probit.**

Applications

Dans les simulations qui suivent, on suppose que la distribution sous $H0$ est la distribution nulle théorique, à savoir uniforme sur $[0,1]$ en ce qui concerne le p-value, et normale de moyenne $\mu_0 = 0$ et $\sigma_0 = 1$ en ce qui concerne les x_i issus de la transformation probit. Cette hypothèse sera discutée par la suite.

Simulations : les simulations consistent en un tirage de n variables aléatoires suivant des distributions normales $\mathcal{N}(0,1)$ et $\mathcal{N}(\mu_1, \sigma_1)$ avec des probabilités *a priori* π_0 et $\pi_1 = 1 - \pi_0$ respectivement. Nous avons fait varier deux paramètres du modèle : μ_1 varie entre -4 et 0 tandis que π_1 varie entre 0 et 0.2. Nous avons fixé μ_0, σ_0 et σ_1 à 0, 1 et 1 respectivement.

Les résultats montrent que l'algorithme EM converge relativement rapidement vers les estimateurs de μ_1, σ_1 au π_1 au maximum de vraisemblance (de l'ordre d'une centaine d'itération, figure 2.20 p. 95). La figure montre également, pour un cas particulier, l'histogramme des x_i ainsi que l'estimation de la fonction densité f_1 qui colle dans ce cas parfaitement à l'observation. On peut par ailleurs remarquer figure 2.21 page 96 que le fait de diminuer la distance entre $H0$ et $H1$ - ce qui revient à rapprocher le paramètre μ_1 de μ_0 - a pour effet d'augmenter la variance des estimateurs des paramètres du modèle ainsi que le temps de convergence de l'algorithme. Cela est naturel puisque moins il y a de différence entre $H0$ et $H1$ et moins l'algorithme a de facilité à distinguer les deux classes du mélange. De la même façon le fait de rapprocher p_1 de 0 augmente la variance de $\widehat{\mu}_1$ et $\widehat{\sigma}_1$ ce qui paraît également assez naturel puisque moins il y a de points dans l'une

des classes et moins l'algorithme a de facilité à distinguer cette classe de l'autre. On peut enfin remarquer que les espérances des estimateurs sont quant à elles, toujours centrées sur la vraie valeur des paramètres.

Données réelles : nous avons appliqué l'estimation du FDR Local sur les données AIM-Scan *genome-wide* concernant la sclérose en plaque. Le jeu de données comprend 114,802 SNPs ; l'ADN de 279 patients et 301 témoins suédois ont été génotypés en utilisant une puce *Affymetrix* 100K. Les p-values associées à chaque marqueur (pv_i) auxquelles nous appliquons la tranformation probit, résultent du test d'Hardy-Weinberg appliqué aux contrôles afin d'identifier les SNPs susceptibles d'être sujets d'erreurs de génotypage (introduction p. 32). Pour appliquer la transformation probit, nous devons préalablement retirer les p-value égales à 1 puisque $\Phi^{-1}(1) = +\infty$. Cela pose naturellement un problème de troncature dans les distributions que nous avons abordé par ailleurs, mais que nous ne développerons pas d'avantage ici.

Nous avons réalisé deux situations : une où la distribution sous $H0$ est connue et où μ_0 et σ_0 sont fixés à 0 et 1 respectivement, et une autre où les paramètres de la classe $H0$ sont estimés par l'algorithme EM. Dans le premier cas, les estimations des paramètres du mélange sont $\widehat{\mu}_1 = -1.36$, $\widehat{\sigma}_1 = 2.1$ et $\widehat{\pi}_1 = 0.069$ (figure 2.22 p. 97). On voit que l'estimation du FDR Local augmente avec la valeur de la p-value jusqu'à la valeur de 0.5 puis diminue. Il s'agit en fait d'un artefact de la méthode dû à la forme gaussienne que l'on impose sur les densités ; comme il y a peu de chances que l'on s'intéresse en pratique aux marqueurs dont la p-value est plus grande que 0.5, on peut les inclure dans les données observées pour estimer la distribution sous $H0$ mais il n'y a pas de raison de chercher à leur donner une interprétation en terme de FDR Local. Dans le cas où l'on estime également les paramètres de la distribution sous $H0$, les résultats sont très similaires (figure 2.23 p. 98). Les paramètres de la classe sous $H1$ sont estimés à $\widehat{\mu}_1 = -1.56$, $\widehat{\sigma}_1 = 2.36$ et $\widehat{\pi}_1 = 0.046$ et ceux de la classe sous $H0$ sont estimés à $\widehat{\mu}_1 = -0.06$, $\widehat{\sigma}_1 = 1.02$ et donc très proches de 0 et de 1 respectivement. Par conséquent l'hypothèse que la distribution sous $H0$ suit une normale $\mathcal{N}(0,1)$ est dans ce cas tout à fait réaliste et l'économie de cette hypothèse ne change pas vraiment les estimations.

Discussion sur le FDR Local

Le modèle paramétrique de mélange gaussien que nous avons présenté est un cadre naturel simple et rapide pour estimer le FDR Local. L'implémentation en R de cette méthode d'estimation est d'ailleurs très succincte :

```
1 x = qnorm(pv)
2 tau1 = runif(length(x))
3 for (i in 1:200){
4   mu1 = sum(x*(1-fdr))/sum((1-fdr))
5   sigma1 = sqrt(sum(x^2*(1-fdr))/sum(1-fdr) - mu1^2)
```

```
    mu0 = sum(x*fdr)/sum(fdr)
    sigma0 = sqrt(sum(x^2*fdr)/sum(fdr) - mu0^2)
    pi0 = mean(fdr)
    fdr = pi0*dnorm(x, mean = mu0, sd = sigma0)/((1-pi0)*
       dnorm(x, mean = mu1, sd = sigma1) + pi0*dnorm(x,
       mean = mu0, sd = sigma0))}
```

Ici, l'algorithme estime les paramètres de la distribution sous $H0$, le nombre d'itérations est fixé à 200 et les valeurs initiales de τ_{1i} sont tirées aléatoirement suivant une loi uniforme $\mathcal{U}(0,1)$, ce qui peut être facilement modifié.

Efron (2004) suggère que la distribution nulle empirique observée dans les données peut s'avérer différente de celle théorique attendue. Cela peut arriver par exemple en raison d'une dépendance entre les observations ou d'une covariable cachée telle que la stratification de population (voir introduction p. 34). Efron propose alors d'ajouter l'estimation des paramètres de f_0 ce qui a pour effet d'augmenter le nombre d'itérations nécessaires à l'algorithme EM pour converger. Dans notre application, l'estimation des paramètres de f_0 ne change pas sensiblement l'estimation du FDR Local du fait que $\widehat{\mu}_0$ et $\widehat{\sigma}_0$ sont proches des valeurs théoriques 0 et 1. Leur estimation n'ajoutant pas à la complexité de l'implémentation et augmentant peu le temps d'exécution sur des données *genome-wide*, il n'y a aucune raison de faire l'économie de l'hypothèse sur les paramètres de la distribution nulle afin de mieux coller à la réalité des données.

De même, concernant la transformation probit appliquée sur les p-values, celle-ci va naturellement être justifiée lorsque les x_i suivent un mélange de distributions gaussiennes. Nous avons vu qu'en pratique, cette hypothèse pouvait fausser l'estimation du FDR Local pour des grandes valeurs de p-values. Cela n'est pas réellement un problème puisque les p-values proches de 1 ne nous intéressent pas en pratique. Auparavant, Allison et al (2002) ont proposé de travailler directement sur les p-values par un mélange de distributions beta et en considérant une uniforme[19] pour la composante nulle du mélange. Cependant, sans plus d'expérience sur la question, nous pensons *a priori* que cette façon de procéder ramène toute l'information intéressante contenue dans le mélange sur 0, ce qui peut affecter la qualité de l'estimation, comparé à des méthodes reposant sur des transformations de type probit ou log par exemple, qui permettent en quelque sorte de "zoomer" sur ce qui se passe au niveau des faibles p-values.

Certaines approches posent des contraintes plus faibles sur les densités nulles et alternatives. Dans notre cas, nous considérons que les marqueurs sont répartis en deux classes : associés et non-associés. Mais il est probable que la classe représentant l'hypothèse alternative soit en réalité un mélange de différentes alternatives arborant différentes forces d'association. Une façon de traiter le problème est d'utiliser un modèle de mélange à plus de deux composantes et d'appliquer un critère de sélection (de type BIC[20]

[19] un cas particulier de la distribution beta

[20] pour *Bayesian Information Criterion*

par exemple) sur le nombre de classes, tel que présenté par Leroux (1992). Une autre approche semi-paramétrique consiste à ne faire aucune hypothèse sur la famille de densités en les intégrant dans le processus d'estimation (Hall 1981, Robin et al 2007). Un tel modèle est extrêmement flexible mais sans contraintes sur la famille de densités, les paramètres du modèle ne sont en pratique pas identifiables. Cette façon d'opérer nécessite donc l'estimation *a priori* des probabilités π_c. Cela peut se faire de différentes manière (voir Storey et Tibshirani 2003 pour un exemple) et l'algorithme EM sur le modèle de mélange paramétrique que nous avons présenté en constitue une. Pour plus de détails, une *review* des différentes méthodes d'estimation du FDR Local et de π_0 est apportée par Dalmasso et al (soumis).

Dans le cadre d'une collaboration autour de l'approche développée par Stéphane Robin et al (2007), nous avons contribué au développement d'un *package* R concernant une estimation semi-paramétrique des modèles de mélange appliquée à la détermination du FDR Local (en développement).

Enfin, les méthodes actuelles d'estimation du FDR et du FDR Local font l'hypothèse que les observations sont indépendantes, hypothèse sous-jacente au modèle de mélange. Dans le cadre des études d'association, on sait que cette hypothèse a peu de chance de se réaliser en raison du déséquilibre de liaison. Mais l'utilisation croissante de marqueurs indépendants à travers la détermination de tagSNPs nous place dans un contexte favorable quant à la validité de cette hypothèse. Par ailleurs, comme nous l'avons constaté dans nos simulations, pour produire des résultats fiables ce type de méthodologie nécessite que la proportion de marqueurs sous $H1$ et la distance entre $H0$ et $H1$ (à savoir la force d'association à la maladie) ne soient pas trop faibles. Cela est pourtant vraisemblablement le cas dans les études d'association sur les maladies complexes où le nombre de marqueurs qu'on l'on s'attend à trouver comme étant associés à la maladie est relativement faible comparé au nombre de marqueurs testés, et où les effets que l'on chercher à mettre en avant sont modestes. Cela n'exclut cependant pas pour autant l'utilisation du FDR Local. Il y a en effet fort à penser qu'en pratique, le choix des marqueurs considérés comme associés à la maladie se fonde plutôt sur une contrainte de moyens limitant le nombre de marqueurs que l'on est capable de post-analyser, que sur le choix d'un seuil statistique. Le FDR Local apporte dans ce contexte une information intéressante pour la validation (bibliographique, expérimentale...) des résultats, plus facile à interpréter pour le généticien que la p-value, et qui permet de jauger la pertinence d'augmenter ou de diminuer le nombre de marqueurs à retenir pour la suite de l'étude, par rapport au nombre que l'on s'était fixé au départ.

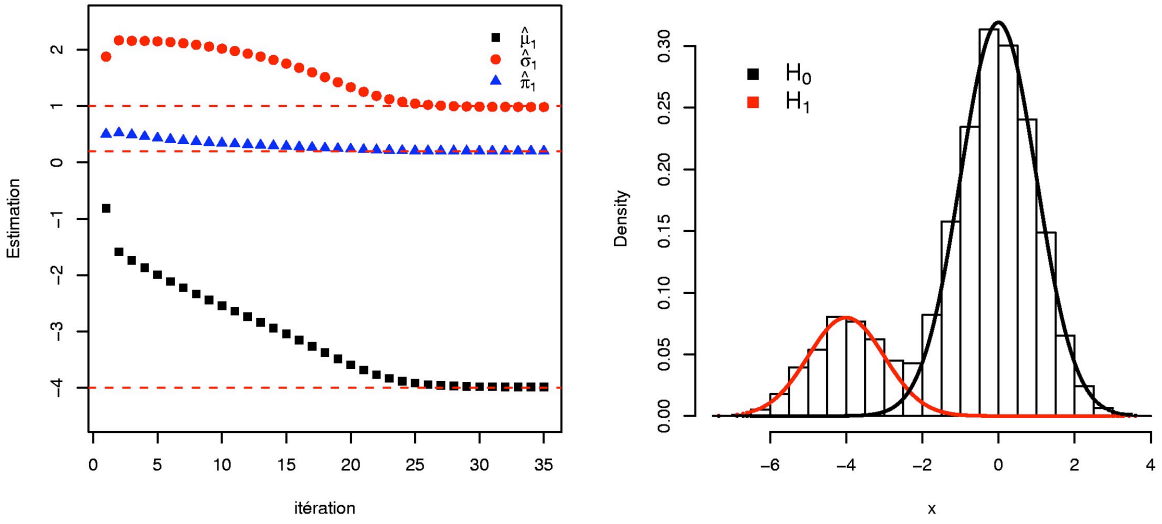


FIG. 2.20 – **Estimation des paramètres d'un modèle de mélange gaussien par algorithme EM : Figure de gauche -** évolution des estimations des paramètres μ_1, σ_1 et π_1 à chaque itération de l'algorithme jusqu'à convergence. **Figure de droite -** histogramme des observations (x_i) et estimation de la densité sous $H1$. La simulation est réalisée pour les paramètres $\mu_1 = -4$, $\sigma_1 = 1$ et $\pi_1 = 0.2$.

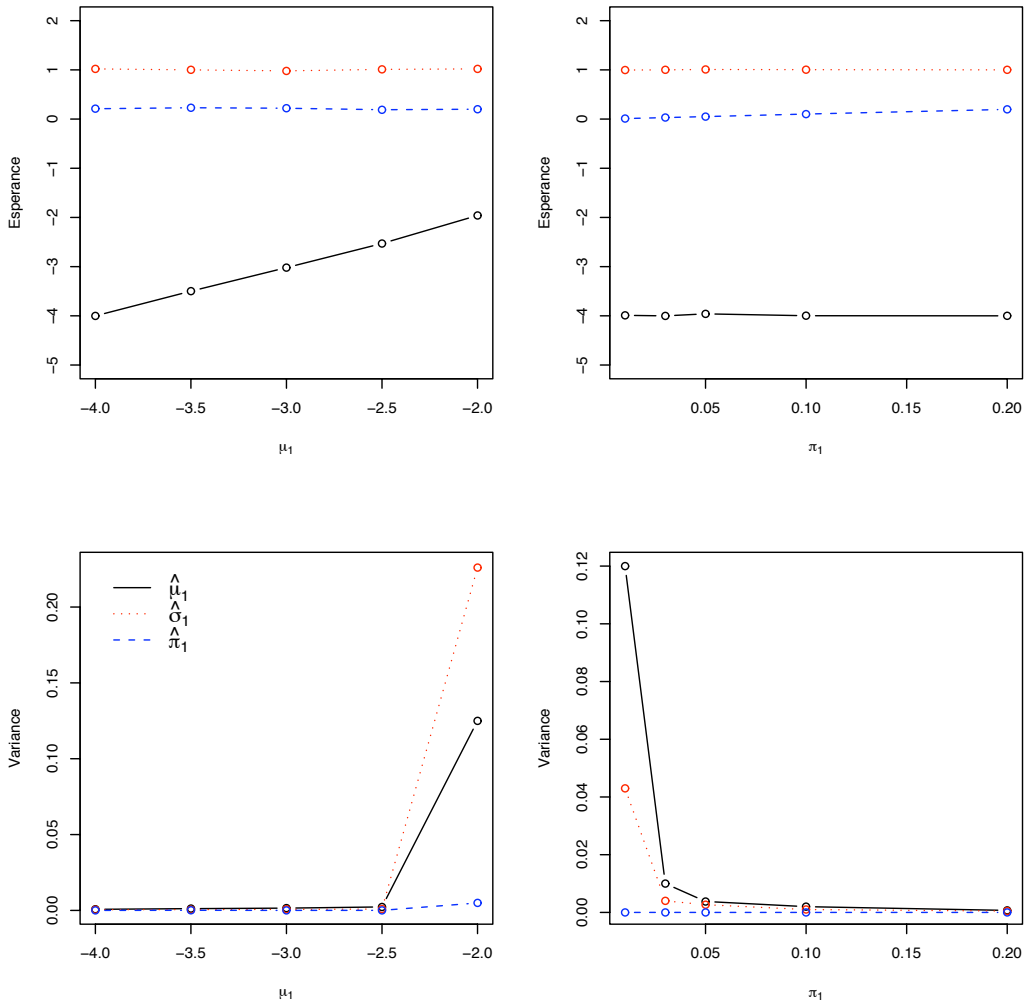


Fig. 2.21 – **Espérance et Variance des estimateurs des paramètres du modèle de mélange gaussien estimés par algorithme EM :** estimés à partir de 1,000 simulations en fixant le paramète σ_1 à 1 et en faisant varier les paramètres μ_1 et π_1.

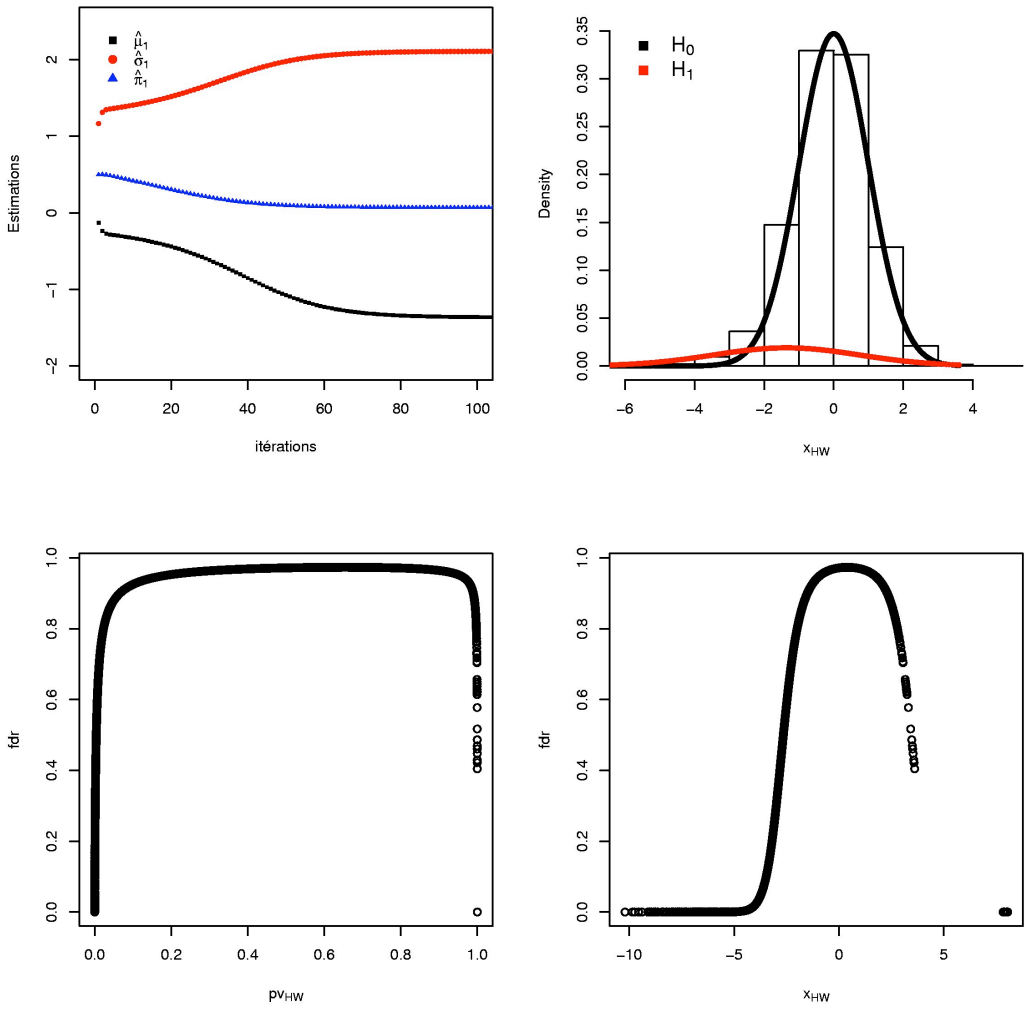


Fig. 2.22 – **Estimation du FDR Local sur les données AIM-SCAN :** appliquée sur les p-values résultantes d'un test d'Hardy-Weinberg appliqué chez les témoins. Les paramètres de la distribution sous $H0$ sont supposés connus : $\mathcal{N}(0,1)$.

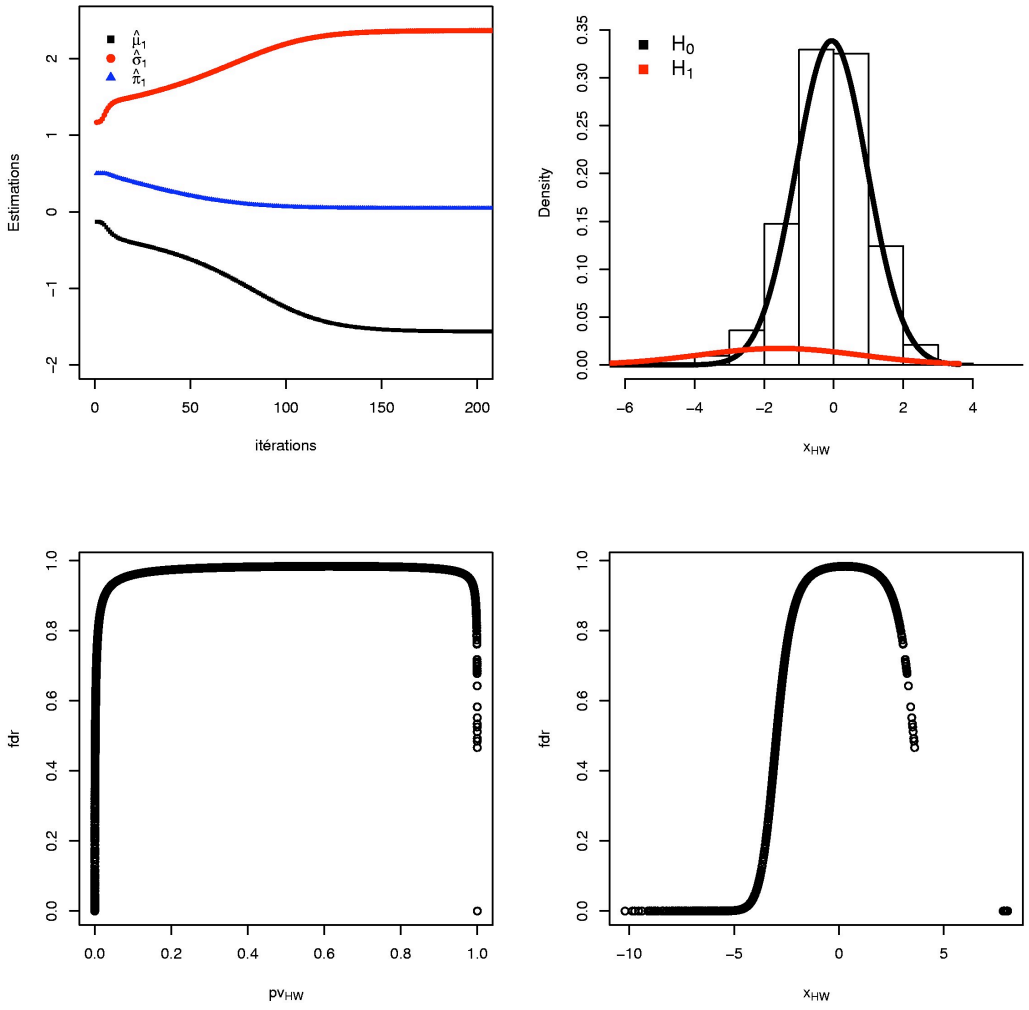


Fig. 2.23 – **Estimation du FDR Local sur les données AIM-SCAN :** appliquée sur les p-values résultantes d'un test d'Hardy-Weinberg appliqué chez les témoins. Les paramètres de la distribution sous $H0$ sont estimés.

2.8 Conclusions

A première vue, mener une analyse simple-marqueur sur une étude d'association cas-témoins peut paraître tout à fait triviale au regard de la simplicité avec laquelle il est possible de considérer les données : il s'agit en fait de tester l'indépendance entre une variable factorielle (le statut) à deux niveaux possibles (affecté, non-affecté) et une autre variable factorielle (le marqueur) à deux ou trois niveaux possibles (l'allèle ou le génotype), de sorte que l'analyse peut se résumer à tester la différence observée de distribution allélique ou génotypique entre les cas et les témoins, en appliquant un simple test du χ^2.

Cependant, le même test peut en fait se décliner sous différentes manières de le réaliser en fonction de la nature même du test[21], du mode d'estimation de la p-value[22] ou encore du modèle d'échantillonnage employé[23]. Toutes ces possibilités vont entraîner des différences sur les résultats obtenus, parfois infimes mais portant parfois à conséquence. En l'occurrence il est bien connu que la qualité d'une estimation asymptotique de la p-value dépend de la réalisation de certaines conditions sur l'échantillon. D'autre part, il est sûrement moins connu que le test exact proposé par Fisher n'est pas strictement équivalent à la version exacte d'un test asymptotique ou empirique réalisé sur la base du score de Pearson. Une première partie importante du travail réalisé dans ce chapitre aura donc été de rappeler les fondements théoriques à partir desquels on a construit les principaux tests d'association en Épidémiologie Génétique.

A partir de ces considérations, nous avons vu qu'il était possible de mettre en place plusieurs tests d'association assez naturels comme le test génotypique, allélique ou encore de tendance. Nous en avons également considéré un quatrième un peu moins intuitif d'un point de vue statistique, fondé sur la déviation par rapport à l'équilibre d'Hardy-Weinberg observé chez les cas.

Décider quelle statistique d'association il convient d'utiliser, est loin d'être évident. Chacune permet de tester l'hypothèse nulle de non-association contre des hypothèses alternatives sensiblement différentes, on se rend rapidement compte que l'efficacité de tel ou tel test par rapport à un autre dépend de la manière dont le locus agit sur la maladie. Cette information étant inconnue, il est impossible de favoriser *a priori* un test plutôt qu'un autre. Dans ce contexte, une stratégie séduisante est de combiner les statistiques ou les tests avec l'attente d'y gagner en terme de puissance. Nous avons parlé de "méta-statistique" afin de désigner ce type d'approche dont la justification est cependant loin d'être claire. De plus, du fait de la dépendance entre les différentes statistiques d'association, un calcul de la p-value autre qu'empirique se révèle délicat.

Afin d'éclairer la décision concernant quelle(s) statistique(s) d'association utiliser, nous avons mis en place une étude de puissance. Cette étude a nécessité dans un premier temps

[21] test de score, de vraisemblance ou de Wald
[22] test asymptotique, empirique ou exact
[23] test conditionnel, non-conditionnel

la définition d'un modèle génétique afin de se placer sous des hypothèses alternatives crédibles. Les principaux paramètres du modèle sont le mode de transmission de la maladie, le déséquilibre de liaison entre le marqueur et locus de susceptibilité dans le cas d'une association indirecte, les proportions alléliques respectives ainsi que le coefficient de consanguinité dans la population. Parmi les différentes méthodes d'estimation de puissance, nous avons considéré les plus couramment utilisées : Monte-Carlo, χ^2 décentré et Delta-Méthode. A ces trois méthodes nous avons ajouté l'utilisation d'une extension de la Delta-Méthode à l'ordre 2 (Forme Quadratique). Nous avons en l'occurrence constaté que malgré une complexité plus importante, la Forme Quadratique permettait une estimation plus rapide que le Monte-Carlo, plus générale que le χ^2 décentré et plus précise que la Delta-Méthode.

Les premières conclusions de notre étude de puissance concernait les tests génotypique et de tendance ainsi que les tests fondés sur la somme et le produit des statistiques correspondantes. Comme attendu, le test génotypique s'avère être meilleur que le test de tendance pour les modes de transmission récessifs, dominant et sur-dominant. Le test de tendance privilégiant une alternative qui va dans le sens d'un mode de transmission additif, il se montre plus puissant pour des modes de transmission additif et multiplicatif. Le choix entre le test génotypique ou le test de tendance n'est donc pas clair. On peut préférer le test génotypique parce qu'il est plus général. Mais on peut aussi penser que les marqueurs indirectement associés à la maladie par déséquilibre de liaison avec un locus de susceptibilité ont plus de chances de présenter une relation à la maladie intermédiaire, même lorsque les locus de susceptibilité pour lesquels ils servent de substitut présentent eux-mêmes des modes de transmission récessif ou dominant. Dans ce contexte, l'utilisation du test de tendance paraît tout à fait raisonnable. La puissance d'une approche fondée sur la combinaison de ces tests ou des statistiques associées se situe presque toujours entre les puissances des tests réalisés seuls. Par conséquent, si l'idée qu'une telle approche apporte un gain de puissance est fausse, elle permet en revanche de proposer une stratégie dont la puissance est intermédiaire, souvent plus proche du meilleur test que du moins bon, et ce, quelle que soit l'hypothèse alternative sous-jacente. Cependant le contrôle du taux d'erreur de type-I nécessite de procéder à ce type de tests avec attention.

Nous avons consacré une section de ce chapitre au test allélique. En faisant l'hypothèse que les allèles sont échantillonnés de manière indépendante dans la population, ce test introduit un biais car cette hypothèse n'est pas réaliste. En pratique, cela induit une déviation des proportions génotypiques observées dans la population combinée par rapport à celles attendues à l'équilibre d'Hardy-Weinberg et une estimation erronée de la p-value. Nous avons étudié l'impact du biais sur données réelles et avons constaté que les conditions dans lesquelles sont généralement réalisées les études *genome-wide* ne sont pas favorables à l'utilisation de ce test. Comme alternative au test allélique biaisé, nous avons proposé un test allélique non-biaisé et exact. Nous avons également mis l'accent sur l'exécution rapide du test. Dans le cadre de notre étude de puissance, nous avons comparé le test allélique non-biaisé au test de tendance qui représente une autre alternative au test allélique biaisé. Les deux tests présentent des valeurs de puissance tout à fait comparables ; le choix entre les deux alternatives repose donc plus sur des considérations pratiques. Pour faciliter la

diffusion de notre test, nous le proposons sous trois versions différentes implémentées par Karl Forner (C++, Perl et R) et disponibles sur le site du laboratoire Statistique et Génome[24].

Une autre section a été consacrée au test d'Hardy-Weinberg. L'utilisation de ce test en tant que test d'association est un peu moins naturelle et provient du fait qu'une association à la maladie entraîne également une déviation chez les cas des proportions génotypiques par rapport à l'équilibre d'Hardy-Weinberg. Ce test permet en l'occurrence de déceler de l'association en présence d'une population de cas uniquement et de préciser la localisation du locus de susceptibilité. Néanmoins, il repose sur des hypothèses fortes. La première et la plus inquiétante est que la population générale doit être à l'équilibre d'Hardy-Weinberg pour contrôler le taux d'erreur de type-I au niveau souhaité. Nous avons vu que dans le cas contraire, ce taux augmente très rapidement avec la déviation par rapport à l'équilibre. Dans un deuxième temps, nous avons également montré que ce test est totalement inapproprié lorsque le mode de transmission sous-jacent est multiplicatif (ou presque) ce qui, comme nous l'avons discuté, a des chances d'être généralement le cas. Dans ce contexte nous ne recommandons donc pas l'utilisation de ce test en tant que test d'association. Nous avons présenté une alternative ($\Delta_{\mathcal{F}}$) qui repose sur la différence entre les coefficients de consanguinité estimés chez les cas et les témoins, et qui permet le contrôle du taux d'erreur de type-I quelle que soit la déviation par rapport à l'équilibre d'Hardy-Weinberg dans la population générale. Ce contrôle se fait néanmoins au coût d'une perte de puissance globale conséquente. L'intérêt d'utiliser l'information apportée par Hardy-Weinberg réside en réalité dans le gain de puissance constaté dans certaines circonstances lorsqu'il est associé aux statistiques d'association plus classiques, du fait qu'il s'appuient sur des aspects différents des données. Plusieurs tests développés récemment reposent sur cette constatation.

Un dernier point important évoqué dans ce chapitre est le choix du seuil de rejet de l'hypothèse nulle lorsque l'on se trouve face au problème du test-multiple. Il est évident que le taux d'erreur de type-I n'est plus une quantité appropriée pour assurer cette décision. Il a été successivement proposé de guider son choix sur le contrôle du *Family-Wise Error-Rate*, souvent considéré comme trop conservatif à un niveau traditionnel de 1% ou 5%, puis du *False Discovery Rate* qui a l'avantage de proposer une estimation du nombre de faux-positifs sur l'ensemble des positifs que l'on obtient à un seuil donné. L'inconvénient du FDR est qu'il considère que tous les positifs partagent la même chance d'être des fausses découvertes, ce qui est naturellement faux et dépend directement de la valeur de la statistique par rapport à la valeur attendue sous l'hypothèse nulle. Pour répondre à ce problème nous avons introduit une quantité proposée récemment : le FDR Local. A cette occasion, nous avons mis en avant une méthode d'estimation qui nous semble être la plus simple et la plus naturelle, reposant sur l'estimation paramétrique d'un modèle de mélange gaussien par algorithme EM, et dont l'implémentation se résume en quelques lignes de code. Cette méthode est rapide et assez flexible pour intégrer si nécessaire l'estimation des paramètres de la distribution des p-values sous l'hypothèse nulle. Elle impose

[24] `http://stat.genopole.cnrs.fr/software/fueatest`

néanmoins un cadre gaussien à deux classes. Dans le but de mieux s'ajuster aux données, des améliorations ont été proposées dans la littérature afin de permettre une estimation des distributions non contrainte à une forme gaussienne et de prendre en compte plus d'une alternative possible. Dans ce contexte et en collaboration avec l'équipe de Stéphane Robin, nous avons participé au développement de l'approche semi-paramétrique qu'il ont proposée. Le *package* R correspondant devrait s'intituler `kerfdr` et être disponible rapidement.

Pour résumer le travail que nous avons effectué sur les approches simple-marqueur, celui-ci s'articule essentiellement autour de trois axes : **(i)** le rappel des fondements théoriques à l'origine des tests d'association, **(ii)** une étude de puissance qui nous a permis de faire le point sur les différentes stratégies possibles et de souligner des problèmes de validité statistique quant à l'utilisation de certains tests et **(iii)** le choix du seuil de rejet de l'hypothèse nulle en considérant le problème du test-multiple. Ce chapitre est cependant loin de couvrir toutes les problématiques liées à ce type d'analyse : les erreurs de génotypage, les valeurs manquantes ainsi que la stratification de population sont autant de thématiques qu'il est important de prendre en compte afin d'assurer la validité et la fiabilité des résultats. Ces différents points ont été évoqués en introduction (p. 28).

Chapitre 3

Approches multi-marqueurs

Il est vraisemblable que la plupart des maladies complexes met en jeu un certain nombre de locus de susceptibilité aux effets modérés, plutôt qu'un seul locus ayant un effet majeur sur la maladie. Par ailleurs ces différents locus sont liés les uns aux autres par des associations alléliques locales générées par le déséquilibre de liaison, et par de possibles effets d'interactions entre locus distants. Dans ce contexte une stratégie d'analyse de type simple-marqueur semble limitée pour élucider l'ensemble des mécanismes impliqués dans les maladies complexes.

Pour répondre à ce problème, plusieurs familles de méthodes d'analyse multi-marqueurs ont été développées et nous en présentons les principales dans la première partie de ce chapitre.

Dans une deuxième partie, nous introduisons l'approche multi-marqueurs que nous avons développée dans le cadre de cette thèse, reposant sur la statistique du Score Local et publiée dans *Statistical Applications in Genetics and Molecular Biology* (2006). Celle-ci consiste à mettre en avant des régions génomiques associées à la maladie s'appuyant sur des accumulations de statistiques d'association élevées le long du génome, générées par le déséquilibre de liaison ou l'agrégation de locus de susceptibilité. L'utilisation du Score Local nous a permis de mettre en place une approche simple, rapide et visiblement plus performante que les approches simple-marqueur traditionnelles. En élevant l'unité de l'analyse du marqueur à une région[1], elle permet de réduire le problème du test-multiple en diminuant le nombre de tests à effectuer. A la suite d'applications sur des données réelles et simulées, nous détaillons les caractéristiques, avantages et inconvénients de cette nouvelle approche, et la replaçons dans le contexte actuelle des analyses multi-marqueurs.

Notes bibliographiques : la rédaction de ce chapitre s'est appuyée en partie sur la lecture de Heidema et al (2006), Robelin (2005) ainsi que Hoh et Ott (2003).

[1] formée par un ensemble de marqueurs contigus

3.1 Introduction

Lorsque l'on s'intéresse aux maladies complexes, une grande difficulté est liée à la prise en compte du rapport qu'entretiennent les marqueurs les uns avec les autres. D'un point de vue local, chaque marqueur est associé par le **déséquilibre de liaison** à ses voisins et éventuellement à un locus de susceptibilité. D'un point de vue plus général, un individu peut être affecté en raison de l'effet joint d'un certain nombre de locus de susceptibilité ; il devient en effet de plus en plus évident que ce type de maladie résulte de l'action d'une combinaison de génotypes présents sur différents locus éventuellement très distants, voire positionnés sur différents chromosomes, que d'un seul gène majeur. On parlera alors d'**interactions génétiques** ou d'**épistasie**.

Dans ce contexte, les approches simple-marqueur qui considèrent uniquement les effets marginaux observés de chaque marqueur sur la maladie, atteignent leurs limites et les statisticiens ont dû se pencher sur le développement de méthodes plus élaborées, capables de tenir compte des relations complexes qui existent entre les locus de susceptibilités et par extension, entre les marqueurs inclus dans le jeu de donnée.

Les premières tentatives d'analyse de plus d'un locus remontent à 70 ans : en 1932, Hogben (1932) suggère que des paires de locus non liés pouvaient être impliquées dans l'apparition d'une maladie. Il a déduit analytiquement les proportions génotypiques attendues pour chaque paire de locus chez les individus atteints à la deuxième génération conditionnellement au phénotype des parents, et les a comparées aux données observées. Cette étude à été la première à considérer plus d'un gène de susceptibilité simultanément. Plus récemment, Gabriel et al (2002) mettent en avant le fait que la transmission visiblement non-mendélienne de la maladie de Hirshchspung est due à l'effet joint de trois gènes positionnés sur trois chromosomes différents. Enfin Martin et al (2002) soulignent une interaction entre le gène *HLA* et le gène *KIR* situés sur le chromosome 6 et le chromosome 9 respectivement ; la combinaison des allèles HLA-BW4-80ILE et KIR-3DS1 est associée à la progression lente du SIDA chez les individus séropositifs. D'autres exemples sont donnés dans la littérature dont un échantillon est proposé par Hoh et Ott (2003).

Le prise en compte de l'effet de plusieurs variables prédictives (ici les marqueurs) et de leurs interactions n'est pas sans poser des problèmes statistiques importants :

(i) De façon analogue à une analyse simple-marqueur, l'importance du nombre de tests que l'on va réaliser peut rapidement atteindre un niveau critique, en particulier si l'on envisage de tester toutes les interactions résultant des combinaisons de marqueurs. Le nombre de tests à réaliser est quadratique avec le nombre de marqueurs pour des interactions d'ordre 2, cubique pour des interactions d'ordre 3, *etc...* ce qui pose à nouveau un problème de **test-multiple** conséquent et soulève également un problème de **temps d'exécution**. A titre illustratif, le traitement de 500,000 marqueurs nécessite d'effectuer au moins 1.25×10^{11} tests rien que pour étudier les interactions d'ordre 2 entre chaque couple de marqueurs. Le test-multiple a été abordé en introduction page 30 et dans le

chapitre 2 page 87.

(ii) Le deuxième problème est lié à la taille modeste des échantillons comparée au grand nombre de marqueurs à tester. On parle alors de **fléau de dimension** (terme introduit par Bellman 1961) qui intervient lorsque le nombre d'observations (les individus dans notre cas) devient trop petit face au nombre de prédicteurs à tester (les marqueurs). Dès lors, certaines valeurs possibles de combinaisons de prédicteurs ne vont pas être observées, particulièrement lorsque l'on va s'intéresser à des ensembles de marqueurs ou des termes d'interactions d'ordre supérieur à 2 et la qualité de l'estimation des paramètres d'un modèle peut en être affecté.

(iii) Le troisième problème est l'analyse de plusieurs marqueurs en présence de données associées par de **déséquilibre de liaison**. La puissance de certaines approches pour identifier les principaux prédicteurs peut en effet diminuer lorsque les prédicteurs testés sont ne sont pas indépendants.

Les approches multi-marqueurs peuvent se répartir en deux grandes familles : les approches dites **multi-points** étudient un ensemble de marqueurs dans le but d'améliorer la détection ou la localisation d'un locus de susceptibilité ; les approches dites **multi-locus** étudient un ensemble de marqueurs dans le but de mettre en évidence plusieurs locus de susceptibilité, positionnés éventuellement sur des chromosomes différents. C'est à cette catégorie que nous nous sommes plus particulièrement intéressés. Les lecteurs néanmoins intéressés par les approches multi-points peuvent se référer à la très large littérature sur le sujet comprenant des méthodes s'appuyant par exemple sur une modélisation du déséquilibre de liaison, sur la reconstruction des haplotypes ou encore sur une interprétation graphique du problème (Stephens et al 2001, Morris et al 2003, Tregouet et al 2004, Schaid 2004, Thomas et Camp 2004, Coulonges et al 2006).

Ce chapitre s'organise autour de deux grands axes. Dans une première section nous introduisons les principales approches multi-locus existantes tout en nous efforçant de mettre en avant leurs principales caractéristiques. Ensuite nous présentons celle que nous avons développée, reposant sur la statistique du Score Local. Le présentation de cette nouvelle approche multi-marqueur passe par une introduction du Score Local (incluant quelques rappels probabilistes, pratiques et algorithmiques), le développement de l'algorithme `LHiSA` dédié à l'utilisation du Score Local dans les études d'association, et son application à des jeux de données réels et simulés. Enfin nous discutons les avantages et inconvénients de cette approche en la replaçant dans le contexte actuel des études multi-marqueurs.

3.2 Approches multi-locus existantes

Régression Logistique

- Régression Logistique : il s'agit d'une approche classique pour modéliser les relations entre des variables prédictives telles que les marqueurs et une variable factorielle à prédire telle que le statut (affecté ou non-affecté). Nous l'avons évoquée dans le chapitre 2, dans le cas d'un seul marqueur (p. 49). L'application de la Régression Logisitique à plusieurs marqueurs est une extension naturelle de l'approche traitant un seul marqueur. Si l'on considère que chacun des n marqueurs est représenté par deux niveaux x_{i1} et x_{i2}, alors le modèle général est à $2n$ degrés de liberté et s'écrit de la manière suivante :

$$\text{logit}(p_{x_1,\ldots,x_n}) = \alpha + \sum_{i=1}^{n} \beta_{i1}x_{i1} + \beta_{i2}x_{i2}.$$

L'on peut également supposer que les effets des allèles sont additifs et que chaque marqueur est représenté par une variable x_i qui prend la valeur 0, 1 ou 2 si le génotype est aa, aA ou AA respectivement. Le modèle est alors à n degrés de liberté et s'écrit sous la forme suivante :

$$\text{logit}(p_{x_1,\ldots,x_n}) = \alpha + \sum_{i=1}^{n} \beta_i x_i.$$

A chaque fois qu'un marqueur est ajouté au modèle, il est possible d'y ajouter autant de termes d'interaction d'ordre 2 qu'il y a de marqueurs déjà présents dans le modèle. Le nombre de termes d'interaction augmente donc exponentiellement avec le nombre de marqueurs, et ce, même en se limitant aux interactions d'ordre 2. Le modèle général s'écrit alors de la façon suivante :

$$\text{logit}(p_{x_1,\ldots,x_n}) = \alpha + \sum_{1}^{n} \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{2}\sum_{l=1}^{2} \gamma_{ijkl}x_{ik}x_{jl}.$$

Les estimations des paramètres ainsi que les tests sur ces paramètres se réalisent de la même façon que dans le cas d'un seul marqueur. Cependant, il est possible de tester la pertinence de l'ajout d'un ou d'un groupe de marqueurs en comparant la vraisemblance du modèle avec la vraisemblance du modèle sans ce(s) marqueur(s). On dit que les deux modèles sont emboîtés et l'on parlera alors de tests sur les modèles emboîtés.

Une telle approche souffre néanmoins de limitations plutôt conséquentes lorsque l'on travaille sur un grand jeu de marqueurs. En outre elle réagit assez mal au problème de dimensions des données évoqué plus haut. Un nombre restreint d'individus comparé au nombre de marqueurs à analyser peut biaiser les estimations des paramètres, augmenter le taux d'erreur de type-I et II (Peduzzi et al 1996) et diminuer la puissance pour détecter

des interactions (Moore and Williams 2002). Un moyen de contourner ce problème est de procéder à une sélection itérative[2] des marqueurs. On parlera de **sélection *forward*** quand on part du modèle nul ($\text{logit}(p) = \alpha$) et que les principaux marqueurs sont ajoutés au modèle en fonction de leur pertinence c'est à dire l'information qu'ils apportent pour expliquer la maladie. Ils sont ajoutés du plus pertinent au moins pertinent avec un critère d'arrêt du type : *plus aucun marqueur ne contribue significativement à améliorer l'explication de la maladie*. A l'inverse une **sélection *backward*** partira du modèle complet incluant tous les marqueurs et soustraira un à un les marqueurs les moins pertinents avec un critère d'arrêt du type : *le retrait d'un prédicteur supplémentaire diminue significativement l'explication de la maladie*. Ces méthodes de sélections, bien que standards, ne sont pas sans présenter quelques lacunes. Par exemple, la sélection *forward* ne considère que les interactions des marqueurs inclus dans le modèle. On peut également se rendre compte que les modèles retenus sont bien souvent sous-optimaux. Des alternatives plus efficaces ont été proposées dans la littérature telles que le *Least Absolute Shrinkage and Selection Operator* (LASSO, Tibshirani 1996).

Par ailleurs, en fonction de la façon dont elle est appliquée, la Régression Logistique peut être très sensible à la dépendance entre les marqueurs qui peut complètement changer les résultats des tests et donc affecter sa puissance. Enfin, il s'agit avant tout d'une approche paramétrique ; elle impose donc des relations fixées et linéaires entre les marqueurs, les valeurs qu'ils peuvent prendre et le statut, ce qui contraint de poser des hypothèses qui peuvent être éloignées de la réalité.

Approches méta-statistiques : sommes de statistiques

Toute une famille d'approches multi-marqueurs combine l'information pertinente portée par un ensemble de marqueurs de façon très simple : par l'intermédiaire de sommes de statistiques d'association simple-marqueurs. On revient ici sur l'idée d'une analyse fondée sur le concept de méta-statistique. La différence avec ce que nous avons vu au chapitre précédent est qu'il ne s'agit plus cette fois de combiner différentes statistiques d'association associées à un marqueur de façon à obtenir une nouvelle statistique d'association pour ce même marqueur, mais de combiner la même statistique d'association appliquée à différents marqueurs de façon à obtenir une statistique d'association pour cet ensemble de marqueurs. L'utilisation de sommes peut être associée à une recherche de simplicité dans la méthodologie utilisée, afin de capter l'information contenue dans des données dont le niveau de complexité peut être très élevé. Nous présentons ici les deux méthodes les plus représentatives de cette famille d'approche : la Fenêtre Glissante et la *Set Association*.

Fenêtre Glissante : il s'agit d'une stratégie qui permet de scanner le génome, en associant à chaque position une statistique résultant de la combinaison des informations

[2] ou *step-wise*

contenues dans une fenêtre de taille fixée. Cette fenêtre est alors déplacée le long du génome, ce qui contraint de travailler sur un jeu de marqueurs contigus. Si n est le nombre de marqueurs et t la taille de la fenêtre, alors celle-ci peut prendre $n-t$ positions possibles notés $F_{1,1+t}, F_{2,2+t}, ..., F_{n-t,n}$. Il existe une grand nombre de manière de traiter l'ensemble de marqueurs contenus dans une fenêtre afin de lui attribuer une statistique d'association $\mathcal{S}'_{i,i+t}$. L'approche la plus classique consiste à sommer les statistiques d'association correspondantes. Cela peut se faire par exemple sur la base des statistiques d'association simple-marqueur $(\mathcal{S}_i)$:

$$\mathcal{S}'_{i,i+t} = \sum_{k=i}^{i+t} \mathcal{S}_k.$$

Cela peut également se faire en utilisant la combaison des p-values de Fisher :

$$\mathcal{S}'_{i,i+t} = -2 \sum_{k=i}^{i+t} \log(p\mathrm{v}_k).$$

Lorsque les marqueurs sont indépendants, cette statistique suis sous l'hypothèse nulle une distribution du χ^2 à t degrés de liberté. Une autre idée consiste à procéder à une reconstruction d'haplotypes au sein de la fenêtre et de proposer une statistique d'association fondée sur une différence de distribution entre les proportions haplotypiques observées chez les cas et chez les témoins. Du fait de la dépendance entre les marqueurs, la significativité de la statistique attachée à chaque fenêtre est généralement évaluée par Monte-Carlo.

L'approche Fenêtre Glissante est sûrement l'une des plus utilisées en pratique dans les études d'association *genome-wide* (Hanson et al 2007 par exemple). En combinant l'information contenue par un ensemble de marqueurs contigus, on espère renforcer la possibilité de détecter un locus de susceptibilité donné. En glissant la fenêtre le long du génome, on s'attend également à mettre en avant un certain nombre de ces locus de susceptibilité. L'approche Fenêtre Glissante peut donc être considérée comme une approche à la fois multi-point et multi-locus qui élève l'unité de base de l'analyse du marqueur à un ensemble de marqueurs contigus de taille fixe. Le principal inconvénient de cette approche réside dans le choix de cette taille de fenêtre. Celui-ci est loin d'être évident si l'on veut coller à une certaine réalité biologique et il n'y a par ailleurs aucune raison pour que la taille de la fenêtre permettant de capter au mieux la dépendance entre les marqueurs soit constante le long du génome.

- *Set Association* : Hoh et Ott (2001) proposent une stratégie qui sélectionne l'ensemble des marqueurs avec le plus de chance d'être associés à la maladie, tout en contrôlant le taux d'erreur de type-I. En pratique une statistique d'association simple-marqueur $(\mathcal{S}_i)$ est calculée pour chaque marqueur. Toutes les statistiques d'association sont ensuite ordonnées $(\mathcal{S}^{(1)} \leqslant ... \leqslant \mathcal{S}^{(n)})$ de façon à ranger les marqueurs (notés $M^{(1)}, ..., M^{(n)}$) du plus associé à la maladie au moins associé. On définit alors N ensembles de marqueurs formés à partir des N meilleurs marqueurs : $\mathrm{Set}^{(1)}, \mathrm{Set}^{(2)}, ..., \mathrm{Set}^{(N)}$ tels que $\mathrm{Set}^{(i+1)} = \mathrm{Set}^{(i)} \cup M^{(i+1)}$ avec $\mathrm{Set}^{(1)} = M^{(1)}$. En pratique, l'ensemble $\mathrm{Set}^{(i)}$ est constitué des i meilleurs marqueurs.

La statistiques d'association correspondant à cette ensemble ($\mathcal{S}'^{(i)}$) résulte alors de la somme des i meilleurs statistiques d'association simple-marqueur :

$$\mathcal{S}'^{(i)} = \sum_{k=1}^{i} \mathcal{S}^{(k)}.$$

La significativité des N ensembles est ensuite estimée par Monte-Carlo. La sélection de l'ensemble le plus pertinent est assurée en considérant celui dont la p-value est la plus petite, appelée $p_{\min}$ par les auteurs et considérée à juste titre comme la statistique et non la p-value de cette sélection. La dernière étape estime donc la significativité de $p_{\min}$ à nouveau par Monte-Carlo.

L'approche *Set Association* a pour objectif de retenir l'ensemble $\text{Set}^{(i)}$ le plus pertinent pour expliquer la maladie. Elle dépend néanmoins du nombre d'ensembles considérés (N) dont le choix n'est guidé par aucun *a priori* biologique ou statistique. Par ailleurs elle considère les marqueurs indépendamment les uns des autres et ne prend donc pas du tout en compte le *pattern* de déséquilibre de liaison qui peut exister dans les données.

De façon générale, les approches fondées sur des sommes de statistiques telles que l'approche Fenêtre Glissante ou l'approche *Set Association* tentent de combiner l'information contenue dans un ensemble de marqueurs contigus pour l'une et distant pour l'autre, de la façon la plus simple qui soit, sans passer par une modélisation formelle, complexe et parfois loin de la réalité des relations qui existe entre les marqueurs à savoir le déséquilibre de liaison et d'éventuels effets d'interaction.

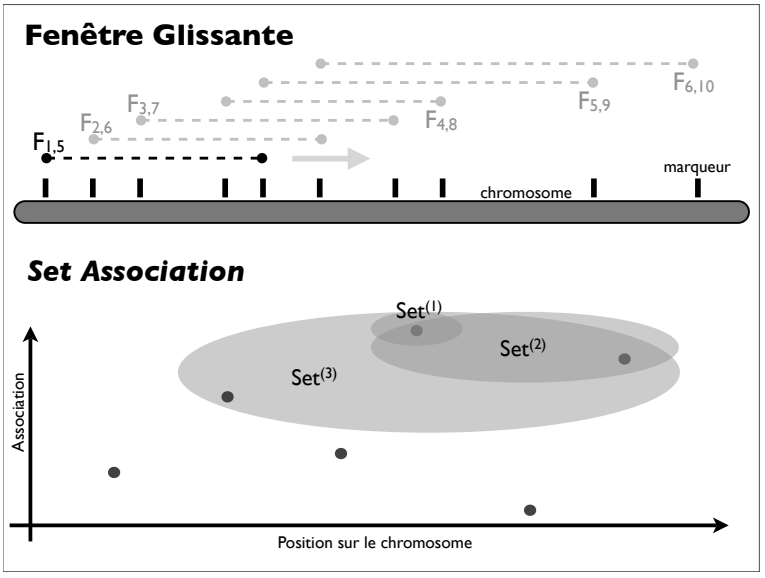


Fig. 3.1 – **Approches par sommes de statistiques.**

Approches combinatoires : *Multiple Dimensionary Reduction*

Les approches combinatoires recherchent les combinaisons de génotypes les plus pertinentes pour expliquer la maladie. En pratique, réaliser le traitement de toutes les combinaisons possibles n'est bien sûr pas réalisable, en particulier lorsque l'on va s'intéresser à des combinaisons mettant en jeu plus de deux marqueurs. Chaque méthode combinatoire apporte donc sa solution quant à l'exploration des combinaisons de génotypes. Nous nous intéressons ici à celle qui nous semble être aujourd'hui la plus connue dans le domaine des études d'association : la *Multiple Dimensionary Reduction* (MDR).

Proposée par Ritchie et al (2001), la MDR est fortement inspirée de la *Combinatorial Partitioning Method* (Nelson et al 2001), une méthode combinatoire développée pour l'analyse de traits quantitatifs. En pratique, la première étape retient les combinaisons de génotypes dites à "haut-risque" pour lesquelles le quotient du nombre de cas par la nombre de témoins possédant cette combinaison dépasse un seuil donné (1 par exemple). Cette opération permet de construire un modèle de classification simple : les individus qui possèdent une des combinaisons de génotypes à haut-risque pour l'ensemble des marqueurs considéré sont prédits comme étant affectés. La pertinence du modèle, c'est à dire la qualité de prédiction, est ensuite évaluée par Validation Croisée[3]. L'ensemble de marqueurs permettant de construire le modèle qui propose la meilleure qualité de prédiction est retenu au final. La significativité des résultats est déterminée par Monte-Carlo en permutant les labels cas et témoins : la p-value est estimée par le nombre de fois qu'un modèle présente une qualité de prédiction supérieure à celle du modèle construit à partir des observations, sur le nombre total de permutations réalisées.

La MDR a été développée afin de proposer une alternative non-paramétrique à la Régression Logistique, permettant de détecter des interactions gène-gène d'ordres élevés, tout en ne posant aucune hypothèse sur le modèle d'interaction sous-jacent. Elle a permis en l'occurrence, sur des données concernant le cancer du sein, de mettre en avant trois gènes impliqués dans le métabolisme de l'oestrogène (Ritchie et al 2001). Cependant en pratique, le temps d'exécution limite sérieusement le nombre de marqueurs ainsi que l'ordre des interactions qu'il est possible d'analyser. De plus, il n'y a aucune indication sur la manière dont la méthode réagit face à l'association allélique résultant du déséquilibre de liaison entre les marqueurs. Pour ces raisons, la MDR apparaît donc plus adaptée aux études gènes-centrés qu'aux études *genome-wide*. Par ailleurs les résultats issus des approches combinatoires ne sont en général pas facilement interprétables, le calcul de la significativité peut être discuté, et dans son cas, la MDR nécessite un nombre égal de cas et de témoins.

[3] mode d'estimation de la fiablité d'un modèle de classification, fondé sur une technique d'échantillonage

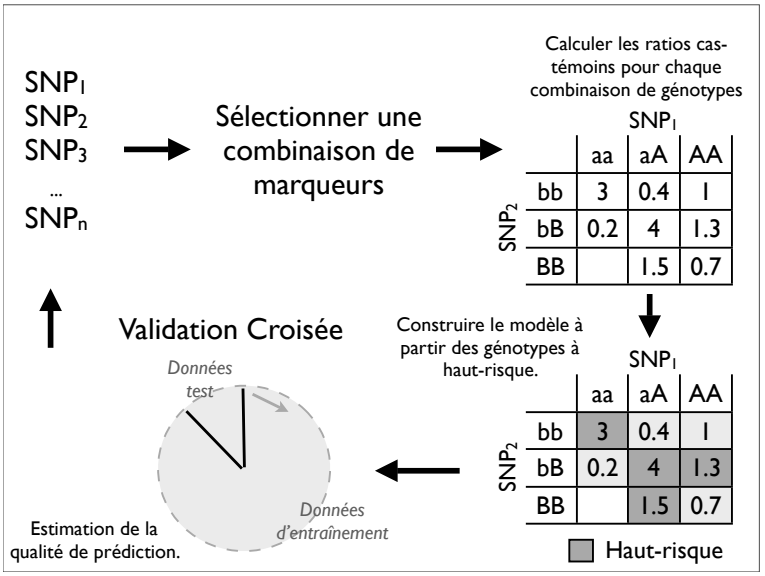


Fig. 3.2 – ***Multi Dimensionary Reduction.***

Approches par partitionnements récursifs : Arbre de Discrimination et Forêt Aléatoire

Les approches par partitionnements récursifs procèdent en partitionnant récursivement l'ensemble des individus en sous-ensemble plus homogènes d'un point de vu de leur statut. Elles sont souvent représentées sous la forme d'un Arbre de Discrimination (Province et al 2001).

Un Arbre de Discrimination est composé d'un noeud racine qui contient le jeu de données complet. Celui-ci est partitionné en deux sous-groupes d'individus par le marqueur qui améliore au mieux l'homogénéité de ces deux sous-groupes. Chacun est lui même partitionné et ainsi de suite jusqu'à ce que l'homogénéité de chaque groupe ne puisse plus être améliorée. Les noeuds terminaux sont appelés les feuilles de l'arbre. Décrit ainsi, un Arbre de Discrimination apparaît équivalent à une sélection de marqueurs *forward* dans une Régression Logistique à la différence essentielle que chaque noeud peut être traité par un marqueur différent.

Une réponse à ce problème consiste à introduire de la flexibilité dans la méthode en ajoutant de l'aléa dans la sélection des marqueurs et en générant un grand nombre d'arbres possibles ; on parle alors de Forêt Aléatoire (Lunetta et al 2004, Bureau et al 2005). Chaque arbre est constitué d'une succession de marqueurs différente. Les marqueurs sont ensuite notés en fonction du nombre de fois qu'il apparaissent dans l'ensemble des arbres et de leur position. Les marqueurs qui apparaissent le plus souvent et/ou placés près de la racine seront considérés comme plus pertinents pour expliquer la maladie que ceux qui apparaissent peu et/ou proches des feuilles.

A l'instar des approches combinatoires, les approches par partitionnements récursifs cherchent à détecter les marqueurs impliqués dans la maladie sans passer par une modélisation formelle des différents effets. Elles ont en l'occurrence pour objectif de mettre en avant des interactions entre marqueurs qui ne présentent *a priori* pas d'effet majeur sur la maladie. Contrairement aux approches combinatoires, elles peuvent être appliquées à un grand nombre de marqueurs. Le vrai problème de ce type d'approche réside dans le choix des critères permettant de retenir au final l'ensemble de marqueurs les plus pertinents, l'interprétation des résultats ainsi que l'estimation d'une significativité statistique.

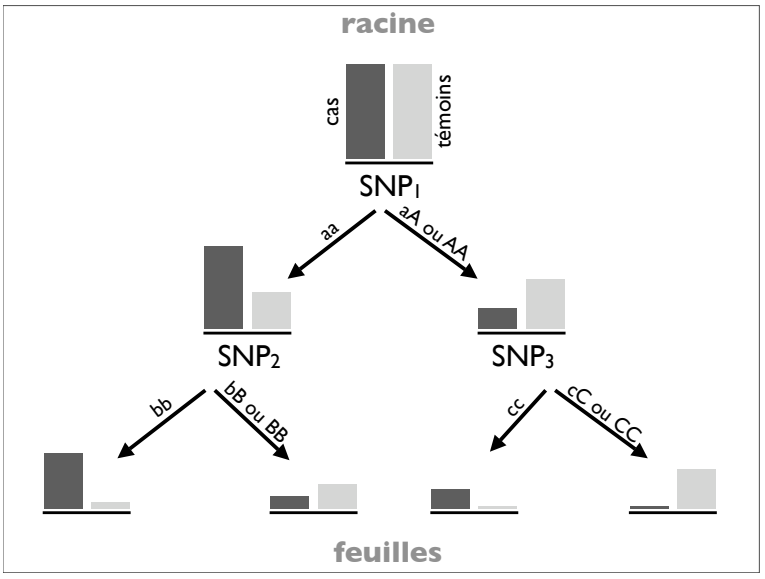


Fig. 3.3 – **Arbre de Discrimination.**

Limites de ces approches et critères d'évaluation

L'ensemble des approches que nous venons d'évoquer permet de souligner la diversité des méthodologies qui ont été développées ces dernières années afin de répondre aux problématiques posées par l'analyse simultanée d'un très grand nombre de marqueurs. Si chaque approche apporte ses solutions, aucune cependant n'est vraiment satisfaisante en tous points. Sans revenir dans le détail sur les avantages et inconvénients de chacune, les approches de type Régression Logistique sont très sensibles au manque d'observations par rapport au nombre de marqueurs à tester ; les approches par sommes de statistiques ne permettent pas en soit de mettre en avant des locus aux effets modestes sur la maladie ou des effets d'interaction ; enfin les approches de classification (combinatoires et par partitionnements récursifs) impliquent une interprétation des résultats et une évaluation de leur significativité statistique pas forcément évidente. De façon plus générale, la plupart de ces approches nécessite de fixer des paramètres dont le choix n'est guidé par aucun *a priori* biologique ou stratégique vraiment clair. De plus elle ne prennent pas vraiment en compte la dépendance locale qui existe entre les marqueurs liée au déséquilibre de liaison.

Afin de résumer ces informations, nous proposons un tableau récapitulatif très personnel des performances de chaque approche (figure 3.4). Nous avons considéré sept critères d'évaluation. Le premier est un critère lié à la **dimension** des données ; il inclut le problème du nombre de marqueurs trop important par rapport au nombre d'observations ainsi que la capacité en temps et en mémoire à traiter de grands jeux de données. Le deuxième critère (**LD**) évalue la capacité à prendre en compte la dépendance locale entre les marqueurs. Le troisième correspond à la **simplicité** de la méthode en terme d'implémentation. Le quatrième note le nombre de paramètres à fixer ainsi que la difficulté liée au **choix de ces paramètres**. Le cinquième évalue la facilité à interpréter les **résultats** ainsi que l'évaluation de leur significativité statistique et le contrôle du taux d'erreur de type-I. Le sixième indique la possibilité de détecter des effets d'**interaction** et le dernier critère celle de mettre en avant des locus qui ont un **effet modeste** sur la maladie.

	RL	FG	SA	MDR	FA
Dimensions	+	+++	+++	+	++
LD	non	pas vraiment	non	non	non
Simplicité	++	+++	+++	+	+
Choix des paramètres	++	++	++	+	+
Résultats	+++	+++	+++	+	+
Interactions	oui	non	non	oui	oui
Effets modestes	oui	non	non	oui	oui

Fig. 3.4 – **Critères d'évaluation :** compare les performances de cinq méthodes multi-marqueurs sur la base de sept critères d'évaluation. Les méthodes sont la Régression Logistique (LR), la Fenêtre Glissante (FG), la *Set Association* (SA), la *Multiple Dimensionary Reduction* (MDR) et la Forêt Aléatoire (FA). Par ailleurs (+), (++) et (+++) représentent une note moyenne, bonne et très bonne respectivement.

3.3 Score Local

Intention de l'approche

Dans les études d'association, on peut s'attendre à une accumulation de hautes valeurs de statistiques d'association le long du génome. De telles accumulations peuvent être dues **(i)** au déséquilibre de liaison entre les marqueurs et le site étiologique ou encore **(ii)** à une agrégation de sites étiologiques dans une même portion du génome. Cette dernière hypothèse est tout à fait réaliste lorsqu'on pense que plusieurs mutations peuvent affecter un même gène ou un *cluster* de gènes voisins impliqués dans une même voix métabolique. Identifier ces accumulations devrait donc aider à identifier des régions génomiques d'intérêt biologique contenant au moins un site étiologique.

De tels accumulations sont faciles à déceler à l'oeil sur une centaine de marqueurs. Cela devient impossible lorsque l'on va passer à l'analyse d'une centaine de milliers de marqueurs. Une façon de les détecter dans les études à grande échelle est d'utiliser le Score Local. Cet outil statistique a été spécifiquement étudié pour identifier des accumulations de valeurs élevées dans une séquence. L'utilisation de cette statistique en biologie n'est pas nouvelle ; elle a été employée avec succès dans l'analyse des séquences biologiques (ADN et protéines) par exemple pour localiser chez les protéines des régions transmembranaires ou hydrophobes, des *DNA-binding regions* ou encore des régions concentrées en charge (Karlin et al 1991, Brendel et al 1992, Karlin et Brendel 1992). Par la suite, l'approche a été étendue à la détection de similarités entre séquences (Altschul et al 1990) : à chaque couple de lettres est associé un score d'autant plus élevé que ces deux lettres se "ressemblent" ; une accumulation de scores élevés désigne donc deux régions très ressemblantes. Pour plus de détails sur les utilisations du Score Local dans l'analyse de séquences biologiques, le lecteur intéressé peut se référer à la *review* proposée par Karlin (2005).

Le Score Local est une statistique déjà bien connue. On se propose ici d'en rappeler la définition, les principaux résultats probabilistes ainsi que les algorithmes mis en jeu. Dans un premier temps nous traitons du Score Local optimal d'une séquence ; nous nous intéressons par la suite à la succesion de Scores Locaux sous-optimaux.

Le Score Local : définition, résultats probabilistes et algorithmes

- Définition : soit $\mathbb{S} = (\mathcal{S}_i)_{i=1,\dots,n}$ une séquence de variables aléatoires. On définit par :

$$H = \max_{1 \leqslant i \leqslant j \leqslant n} \sum_{k=i}^{j} \mathcal{S}_k,$$

le Score Local attribué à $\mathbb{S}$. En pratique, le Score Local est défini comme la valeur de la sous-séquence (ou région) qui présente la valeur maximale pour les sommes des $\mathcal{S}_i$. Cette

valeur est maximale dans le sens où l'on ne peut augmenter ni diminuer la taille de la région sans affecter la valeur du Score Local. De façon plus formelle, Ruzzo et Tompa (1999) apportent la définition suivante :

Soit $\mathbb{S}$ une séquence non vide. Une sous-séquence $\mathbb{I}$ est de score maximal dans $\mathbb{S}$ si, et seulement si **(i)** toutes les sous-séquences de $\mathbb{I}$ ont un score plus faible et **(ii)** aucune sur-séquence de $\mathbb{I}$ contenue dans $\mathbb{S}$ ne satisfait la condition précédente.

La région attachée au Score Local est la "meilleur région" ou encore appelée "*maximal scoring subsequence*", "*locally optimal subsequence*", "*maximum sum interval*" et "*local highest-scoring segment*" dans la littérature.

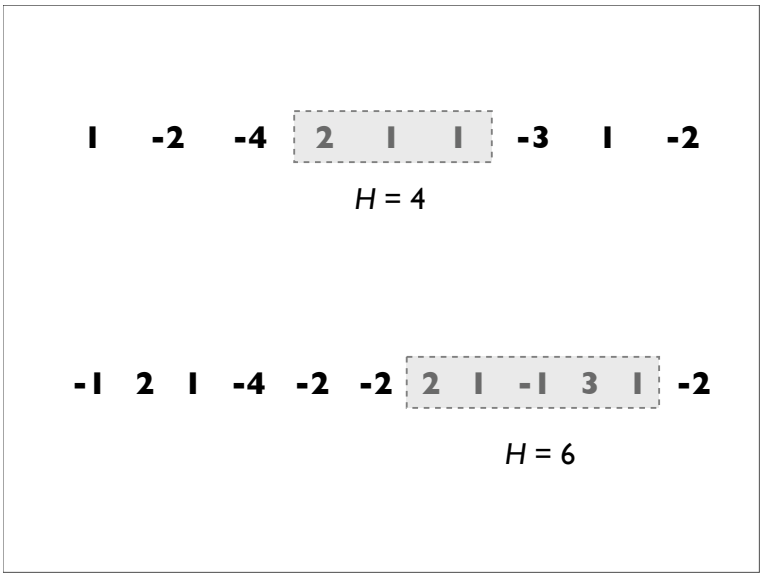


Fig. 3.5 – **Score Local maximal :** exemples.

- Résultats probabilistes : en se référant à la théorie des valeurs extrêmes et au travail réalisé par Iglehart (1972), Karlin et Dembo (1992) ont montré que dans l'hypothèse où les $\mathcal{S}_i$ sont indépendants et identiquement distribués (iid), où $\mathbb{E}(\mathcal{S}_i) < 0$ et où n est suffisamment grand, la distribution de H peut être approchée par une distribution de Gumbel :

$$\lim_{n\to+\infty} \mathbb{P}\left(H \geqslant \frac{\log n}{\lambda} + h\right) = 1 - \exp(-Ke^{-\lambda h}).$$

Les paramètres α et β de la loi de Gumbel tels qu'il sont classiquement donnés dans la littérature sont dans ce cas $\alpha = \frac{\log n + \log K}{\lambda}$ et $\beta = \frac{1}{\lambda}$. La distribution du Score Local dépend donc de la taille de la séquence n ainsi que de deux constantes de normalisation K et λ. On peut exprimer ce résultat plus simplement en considérant le Score Local normalisé $H' = \lambda H - \log(nK)$:

$$\lim_{n\to+\infty} \mathbb{P}\left(H' \geqslant h\right) = 1 - \exp\left(-e^{-h}\right).$$

K et λ dépendent de la distribution de $\mathcal{S}_i$ et sont rarement faciles à calculer en pratique. Le lecteur intéressé trouvera en annexe (p. 196) la démonstration permettant d'obtenir la loi du Score Local ainsi qu'une précision sur l'estimation de K et λ.

On saisit par ailleurs assez bien qu'en pratique l'espérance des $\mathcal{S}_i$ doit être négative, autrement la meilleure région peut facilement recouvrir l'ensemble de la séquence, ce qui n'est naturellement pas le but recherché.

- Algorithmes : il s'agit ici de présenter comment déterminer efficacement le Score Local ainsi que la région i, j correspondante. Une approche naïve consiste à calculer et comparer toutes les valeurs du score ($\sum_i^j \mathcal{S}^k$) pour toutes les tailles de régions possibles. On peut noter qu'il n'y a pas nécessairement unicité de la solution. Certains préconisent de favoriser la région la plus courte ce qui limite le nombre de solutions mais ne résout pas le problème. Dans ce cas, une région est choisie arbitrairement parmi toutes les régions qui réalisent le Score Local optimal :

Algorithme 1
1 : Pour tous les i et j tels que $i \leqslant j \in [1, n]$, calculer $H_{ij} = \sum_i^j \mathcal{S}_k$.
2 : calculer $H = \max(H_{ij})$.

Une telle approche est **cubique** en temps d'exécution avec la taille de la séquence ($\mathrm{O}(n^3)$). Une autre façon de procéder consiste à considérer la marche aléatoire $\mathbb{M} = (M_i)_{i=1,\ldots,n}$ avec $M_i = M_{i-1} + \mathcal{S}_i$ et $M_1 = \max(0, \mathcal{S}_1)$. Le Score Local est alors déterminé par la plus grande différence de marche $M_{ij} = M_j - M_i$:

Algorithme 2
1 : calculer $M_1 = \max(0, \mathcal{S}_1)$.
2 : Pour tout $i \in [2, n]$, calculer $M_i = M_{i-1} + \mathcal{S}_i$.
3 : Pour tous les i et j tels que $i \leqslant j \in [1, n]$, calculer $M_{ij} = M_j - M_i$.
4 : calculer $H = \max(M_{ij})$.

Cette approche est **quadratique** avec la taille de la séquence ($\mathrm{O}(n^2)$). Une dernière façon réellement plus efficace de traiter la séquence, est de ramener toutes les différences de marches négatives à 0 en considérant le processus $\mathbb{H} = (H_i)_{i=1,\ldots,n}$ avec $H_i = \max(0, H_{i-1} + \mathcal{S}_i)$ et $H_1 = \max(0, \mathcal{S}_1)$ (Bates and Constable 1985) :

Algorithme 3
1 : calculer $H_1 = \max(0, \mathcal{S}_1)$.
2 : Pour tout $i \in [2, n]$, calculer $H_i = \max(0, H_{i-1} + \mathcal{S}_i)$.
3 : calculer $H = \max(H_i)$.

De cette manière, trouver le Score Local d'une séquence revient à trouver le maximum du processus $\mathbb{H}$, ce qui est **linéaire** avec la taille de la séquence ($\mathrm{O}(n)$) et non plus quadratique.

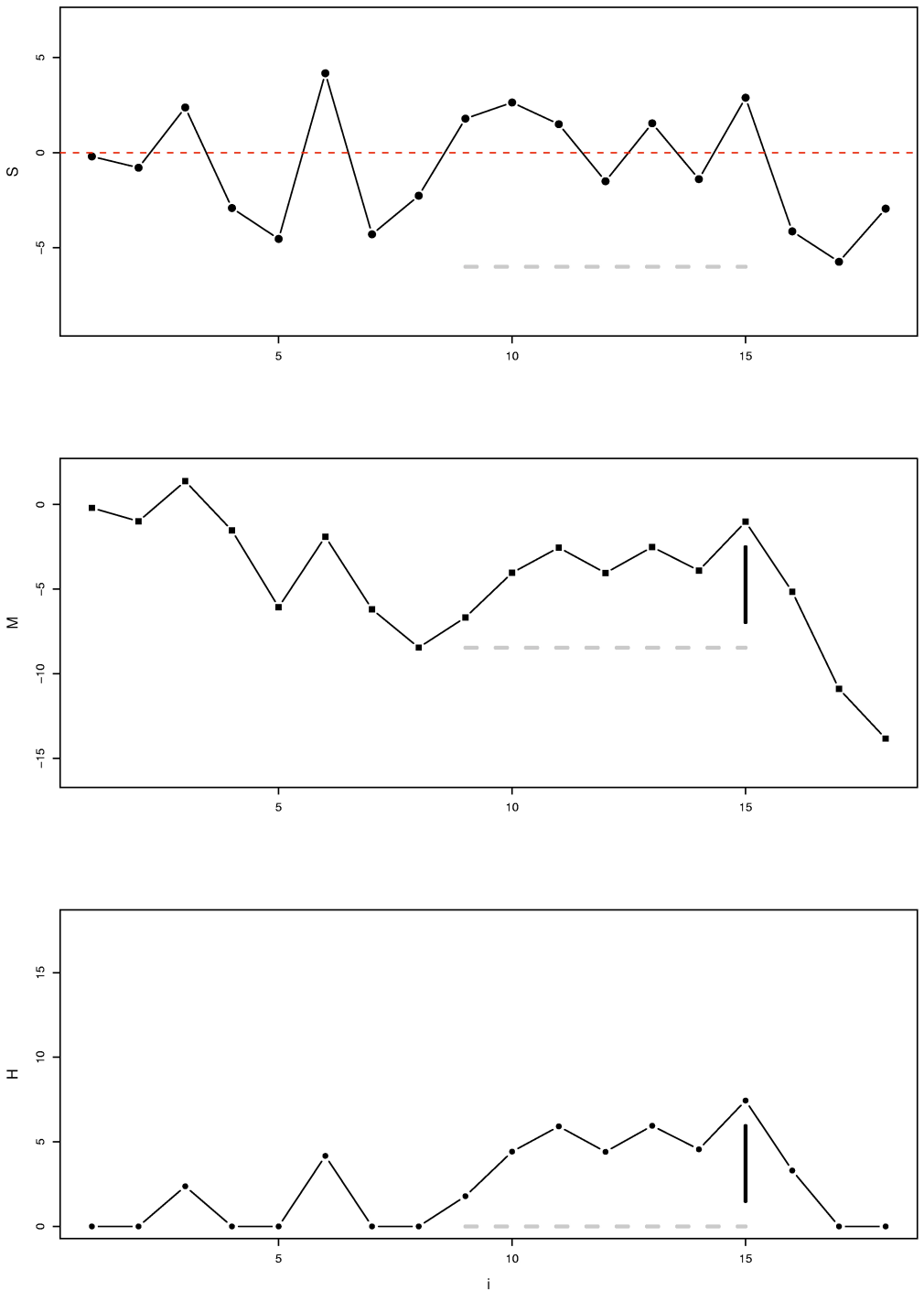


Fig. 3.6 – **Illustration des trois algorithmes de recherche du Score Local optimal :** $\mathbb{X}$ est la séquence sur laquelle on travaille ; H est alors égal à $\max \sum_i^j X_k$. On définit par $\mathbb{M}$ la marche aléatoire associée à $\mathbb{X}$; H est alors égal à la plus grande différence de marche. Enfin, le processus $\mathbb{H}$ ramène toutes les différences de marche négatives à 0. Ainsi, H est égal à la valeur maximal de $\mathbb{H}$.

Le $k$ème Score Local : définition, résultats probabilistes et algorithmes

La définition du Score Local que nous avons introduite se restreint à la recherche de la meilleure région ; cependant les phénomènes biologiques que l'on cherche à déceler avec cette statistique peuvent se répéter plusieurs fois le long de la séquence puisque l'on s'attend, dans notre cas, à ce qu'il existe plusieurs régions d'intérêt. Par conséquent, en plus du Score Local maximal, soit le 1er Score Local ($H^{(1)}$), on va s'intéresser à la recherche du 2ème, 3ème jusqu'au $k$ème Score Local ($H^{(2)}, ..., H^{(k)}$).

- Définition : au sens strict de la définition que nous avons fournie précédemment, la 2ème meilleure région a de grandes chances de recouvrir en partie la 1ère et d'identifier plus ou moins la même portion de génome, ce qui n'apporte aucune information supplémentaire. Par conséquent l'on définit le 2ème Score Local comme le Score Local de la séquence disjointe de la meilleure région, et en toute généralité, le $k$ème Score Local comme étant celui de la séquence disjointe des $k-1$ meilleures régions. L'ensemble des statistiques ordonnées $H^{(1)} \geqslant ... \geqslant H^{(k)}$ correspond alors aux k premiers Scores Locaux de la séquences associés aux k premières meilleures régions disjointes.

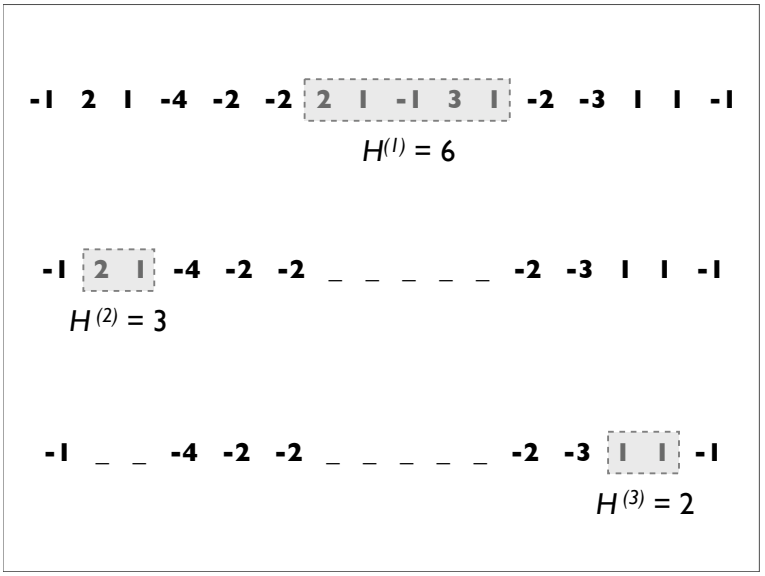


Fig. 3.7 – **Les k meilleurs Score Locaux :** un exemple.

- Résultats probabilistes : Les k premiers Score Locaux $H^{(1)}, ..., H^{(k)}$ peuvent s'écrire sous la forme normalisée $H'^{(1)}, ..., H'^{(k)}$ avec $H'^{(i)} = \lambda H^{(i)} - \log(nK)$ où λ et K sont les constantes de normalisation définies précédemment. A partir de cette formulation, Karlin

et Altschul (1993) proposent une généralisation de la formule pour le $k$ème Score Local :

$$\lim_{n\to+\infty} \mathbb{P}\left(H'^{(k)} \geqslant h'\right) = 1 - \exp\left(-e^{-h'}\right) \times \sum_{i=0}^{k-1} \frac{e^{-ih'}}{i!}.$$

Les auteurs fournissent également une approximation asymptotique de la fonction densité (f_i) pour la somme des i premiers Scores Locaux normalisés, $T'^{(i)} = H'^{(1)} + ... + H'^{(i)}$:

$$\lim_{n\to+\infty} f_i(t') = \frac{e^{-t'}}{i!(i-2)!} \int_0^{+\infty} y^{i-2} \exp\left(-e^{(y-t')/i}\right) dy.$$

Algorithmes : la méthode la plus naturelle et la plus couramment utilisée pour trouver les Scores Locaux consiste à appliquer itérativement l'algorithme de recherche du Score Local optimal en retirant de la séquence les régions attachées aux Scores Locaux déjà identifiés : on trouve le 1er Score Local, on retire la région associée de la séquence, on trouve le 2ème et on réitère jusqu'à ce qu'il n'y ait plus de Scores Locaux positifs dans la séquence.

Algorithme 4
1 : calculer $H^{(1)}$.
2 : retirer la région correspondante de la séquence.
3 : $i = 1$.
4 : tant que $H^{(i)} > 0$:
> $i = i + 1$.
> calculer $H^{(i)}$.
> retirer la région correspondante de la séquence.

Dans le pire des cas, la complexité de l'algorithme est quadratique avec la taille de la séquence (n^2). En pratique une complexité de l'ordre de $n \log n$ semble plus réaliste (annexe p. 196). Au regard de la taille des séquences que l'on traite aujourd'hui, on préfère utiliser l'algorithme suivant, bien plus efficace.

Ruzzo and Tompa (1999) proposent un algorithme de recherche de tous les Scores Locaux positifs présents dans une séquence, d'une complexité linéaire avec la taille de la séquence. La séquence est lue de gauche à droite et à chaque itération, une liste de régions disjointes notées $\mathbb{I}_1, ..., \mathbb{I}_{k-1}$ évolue. A chaque $\mathbb{I}_j$ est associé L_j la somme cumulée de tous les scores à gauche de $\mathbb{I}_j$ et R_j, la somme cumulée de tous les scores depuis le début de la séquence jusqu'à la fin de $\mathbb{I}_j$. Au début de l'algorithme, le liste est vide. Les scores non-positifs ne sont pas traités. Seuls les scores positifs sont inclus dans une nouvelle région $\mathbb{I}_k$ puis traités de la façon suivante :

Algorithme 5
1 : la liste est parcourue de droite à gauche jusqu'à ce que $L_j < L_k$.
2 : s'il n'y a pas un tel j, alors la région $\mathbb{I}_k$ est ajoutée en fin de liste.
3 : s'il existe un tel j et que $R_j \geqslant R_k$, la région $\mathbb{I}_k$ est ajoutée en fin de liste.
4 : dans le cas contraire, la région $\mathbb{I}_k$ est étendue sur la gauche jusqu'à ce qu'elle contienne le segment $\mathbb{I}_j$, les segments numérotés de j à $k-1$ sont éliminés de la liste et le segment $\mathbb{I}_k$ prend la place de $\mathbb{I}_j$ et cette nouvelle région subit à nouveau le traitement depuis l'étape 1.

Au final, la valeur du Score Local d'une région $\mathbb{I}_j$ est donnée par $R_j - L_j$ et la liste des Scores Locaux donnée par l'algorithme n'est pas triée suivant les valeurs de Score Local décroissantes.

- Exemple du déroulement de l'algorithme 5 : déroulons l'algorithme de Ruzzo et Tompa sur l'exemple proposé page 118 et résumé dans le tableau suivant :

index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\mathbb{S}$	-1	2	1	-4	-2	-2	2	1	-1	3	1	-2	-3	1	1	-1
$\mathbb{M}$	-1	1	2	-2	-4	-6	-4	-3	-4	-1	0	-2	-5	-4	-3	-4
$\mathbb{H}$	0	2	3	0	0	0	2	3	2	5	6	4	1	2	3	2

Dans cet exemple il est simple d'identifier à "l'oeil" les trois Score Locaux tels qu'ils sont montrés sur la figure 3.7 p. 118. Les deuxième, troisième et quatrième lignes du tableau correspondent respectivement à la séquence $\mathbb{S}$, à la somme cumulée $\mathbb{M}$ et au processus $\mathbb{H}$ tels qu'ils ont été définis précédemment (figure 3.6 p. 117). Déroulons maintenant l'algorithme. A chaque étape nous indiquerons la liste des régions, chacune des régions $\mathbb{I}_i$ étant décrite par sa position en termes d'index dans la séquence, la somme cumulée au début du segment L_i et la somme cumulée à la fin R_i.

On lit le premier nombre de $\mathbb{S}$, $\mathcal{S}_1 = -1$. Il est négatif donc il n'est pas traité.

On lit $\mathcal{S}_2 = 2$:

$$\mathbb{I}_1 \quad 2 \quad L_1 = -1 \quad R_1 = 1$$

On lit $\mathcal{S}_3 = 1$:

$$\begin{array}{llll} \mathbb{I}_1 & 2 & L_1 = -1 & R_1 = 1 \\ \mathbb{I}_2 & 3 & L_2 = 1 & R_2 = 2 \end{array}$$

$L_1 < L_2$ et $M_2 > M_1$ donc les deux régions sont fusionnées :

$$\mathbb{I}_1 \quad \text{2-3} \quad L_1 = -1 \quad R_1 = 2$$

On passe au $\mathcal{S}_i$ positif suivant. On lit $\mathcal{S}_7 = 2$:

$$\begin{array}{cccc} \mathbb{I}_1 & \text{2-3} & L_1 = -1 & R_1 = 2 \\ \mathbb{I}_2 & 7 & L_2 = -6 & R_2 = 0 \end{array}$$

$L_2 < L_1$ donc les deux régions ne fusionnent pas. On lit $\mathcal{S}_8 = 1$. $L_2 < L_3$ et $R_3 > R_2$ donc les deux régions sont fusionnées et il en est de même jusqu'à $\mathcal{S}_{11} = 1$:

$$\begin{array}{cccc} \mathbb{I}_1 & 2 & L_1 = -1 & R_1 = 1 \\ \mathbb{I}_2 & \text{7-11} & L_2 = -6 & R_2 = 0 \end{array}$$

On passe au $\mathcal{S}_i$ positif suivant. On lit $\mathcal{S}_{14} = 1$:

$$\begin{array}{cccc} \mathbb{I}_1 & \text{2-3} & L_1 = -1 & R_1 = 2 \\ \mathbb{I}_2 & \text{7-11} & L_2 = -6 & R_2 = 0 \\ \mathbb{I}_3 & 14 & L_3 = -4 & R_3 = -3 \end{array}$$

$L_2 < L_3$ mais $R_3 < R_2$ donc les deux régions ne fusionnent pas. On lit $\mathcal{S}_{15} = 1$. $L_3 < L_4$ et $R_4 > R_3$ donc les deux régions sont fusionnées. Le dernier élément de la séquence $\mathbb{S}$ est négatif donc il marque la fin de l'algorithme. Au final, nous avons identifié les trois régions dont le Score Local associé à chaque région $\mathbb{I}_i$ est donné par la différence $R_i - L_i$. On obtient alors la liste non-ordonnée :

$$\begin{array}{cc} \mathbb{I}_1 & \text{2-3} \quad H_1 = R_1 - L_1 = 3 \\ \mathbb{I}_2 & \text{7-11} \quad H_2 = R_2 - L_2 = 6 \\ \mathbb{I}_3 & \text{14-15} \quad H_3 = R_3 - L_3 = 2 \end{array}$$

Une fois ordonnée, la liste des régions est alors donnée par :

$$\begin{array}{cc} \mathbb{I}^{(1)} & \text{7-11} \quad H^{(1)} = R_2 - L_2 = 6 \\ \mathbb{I}^{(2)} & \text{2-3} \quad H^{(2)} = R_1 - L_1 = 3 \\ \mathbb{I}^{(3)} & \text{14-15} \quad H^{(3)} = R_3 - L_3 = 2 \end{array}$$

3.4 Algorithme LHiSA

L'ensemble des connaissances sur le Score Local réunies dans la section précédente a été utilisé pour produire un algorithme en quatre étapes (appelé `LHiSA` pour *Local High-scoring Segments for Association*) adapté aux particularités des études d'association et permettant de mettre en avant des régions du génome statistiquement associées à la maladie considérée. Une première étape attribue à chaque marqueur une statistique d'association pour constituer le signal d'association ; une deuxième étape identifie les Scores Locaux contenus dans le signal d'association ; une troisième étape propose un sous-ensemble de régions potentiellement intéressantes et une dernière étape évalue la p-value associée à cette sélection.

Par ailleurs, l'hypothèse nulle testée ici est $H0$: *il n'y a pas de région associée à la maladie* contre l'hypothèse alternative $H1$: *il y a au moins une région associée à la maladie.*

Étape 1 : générer le signal d'association

Dans un premier temps, on attribue à chaque marqueur une statistique d'association $(\mathcal{S}_i)$ afin de constituer le signal d'association à la base de la méthode : $\mathbb{S} = (\mathcal{S}_i)_{i=1\ldots n}$. La statistique d'association en question peut être n'importe quelle statistique classique d'association (X_G, X_T, X_A,..., voir chapitre 2 p. 39). On peut également considérer les p-values d'association en appliquant une transformation de manière à ce qu'une forte valeur de $\mathcal{S}_i$ implique un haut degré d'association. On peut par exemple appliquée la transformation logarithmique : $-\log_{10}(p\text{v})$.

Comme nous l'avons indiqué précédemment, une contrainte de l'approche par Score Local est que la séquence sur laquelle il s'applique doit être négative en espérance : $\mathbb{E}(\mathcal{S}_i) < 0$. Cela n'est généralement pas le cas, en particulier si l'on considère des statistiques d'association de type χ^2 qui sont positives. Il est alors nécessaire de diminuer le signal d'une constante δ afin d'obtenir un signal d'association $\mathbb{S}' = (\mathcal{S}'_i)_{i=1\ldots n}$ avec $\mathcal{S}'_i = \mathcal{S}_i - \delta$ tel que $\mathbb{E}(\mathcal{S}'_i) < 0$.

Le paramètre δ est le seul à fixer. Il est facilement interprétable puisqu'il correspond au seuil simple-marqueur au dessus duquel on considère qu'un marqueur est potentiellement intéressant et doit positivement contribuer au Score Local, et en dessous duquel il doit le pénaliser. Lui choisir la valeur qui correspond au traditionnel niveau 5% pour la statistique considérée est tout à fait raisonnable ; comme les effets que l'on veut détecter sont vraisemblablement modestes, on peut également considérer d'être un peu plus laxiste sur la valeur de δ et de monter au niveau 10%.

Par exemple, si $\mathcal{S}_i$ est la statistique de Pearson appliquée à la table de contingence

génotypique, que l'on considère raisonnablement que cette statistique suis un χ^2 à 2 degrés de liberté, et que l'on fixe δ à la valeur de cette statistique correspondant au niveau 5%, δ est alors égal à 5.99. Si dans un autre cas $\mathcal{S}_i$ est construite à partir des p-values en appliquant préalablement une transformation logarithmique $-\log_{10}(p\text{v})$, δ est alors égal à $-\log(0.05)$.

L'impact de δ sur les résultats est un aspect important de l'étude des performances de notre méthode qui est discuté par la suite.

Étape 2 : identifier les Score Locaux

Une fois le signal d'association $\mathbb{S}$' généré à partir des données, le but est d'identifier tous les Score Locaux positifs $(H^{(1)}, ..., H^{(k)})$ présents dans le signal ainsi que les régions correspondantes $(\mathbb{I}^{(1)}, ..., \mathbb{I}^{(k)})$. Les algorithmes mis en jeu à cette étape ont été décris en détail dans la section précédente.

Étape 3 : proposer un ensemble de régions

A notre connaissance, seul l'article de Karlin et Altschul (1993) traite dans la littérature du problème du choix des régions à retenir en proposant quelques pistes à suivre sans vraiment proposer une solution aboutie. L'idée est de considérer que parmi les k Scores Locaux identifiés, on va retenir les r premiers. Pour cela les Scores Locaux $H^{(1)}, ..., H^{(k)}$ sont combinés dans une nouvelle somme de statistiques : $T^{(i)} = \sum_{j=1}^{i} H^{(j)}$. En pratique, $T^{(1)}$ représente la sélection de la première région, $T^{(2)}$ la sélection des 2 premières régions et $T^{(k)}$ celle des k régions. Les p-values $(p_T^{(1)}, ..., p_T^{(k)})$ correspondantes à chaque sélection sont estimées en utilisant les résultats probabilistes donnés dans la section précédente ou par Monte-Carlo en obtenant la distribution empirique sous $H0$ de chaque $T^{(i)}$.

Karlin et Altshul considèrent que l'ajout d'une région est biologiquement intéressant du moment qu'il ne diminue pas la significativité de la sélection $(p_T^{(i+1)} \leqslant p_T^{(i)})$. Par conséquent, r correspond aux premières régions qui n'induisent pas une augmentation de la p-value au moment de leur inclusion dans la sélection :

$$r = \text{argmin}_i(p_T^{(i+1)} > p_T^{(i)}).$$

Cette façon de procéder constitue une solution au problème ; il en existe d'autres que nous discuterons et mettrons en perspectives de ce travail.

Étape 4 : calculer la p-value globale

Si Karlin et Altshul (1993) proposent une stratégie qui sélectionne les r régions que l'on va retenir, les auteurs ne sont en revanche pas clairs sur la p-value résultante de cette sélection. Ils considèrent que la p-value globale (p_G) de l'approche est $p_{\min}^{\text{obs}} = p_T^{(r)}$. Or il est évident que si $p_{\min}^{\text{obs}}$ est la p-value associée au fait de sélectionner r régions, elle ne prend absolument pas en compte le fait que r peut varier et donc, de façon plus globale, ne prend pas du tout en compte la procédure de sélection proposée à l'étape 3. Formellement, $p_{\min}^{\text{obs}}$ est la statistique et non la p-value associée à la sélection. Il est donc nécessaire d'estimer p_G par Monte-Carlo en réitérant B fois les étapes 1, 2 et 3, en permutant à chaque simulation (i) les labels cas-témoins et en calculant le $p_{\min}^{(i)}$ correspondant. Cela permet d'obtenir la distribution empirique de $p_{\min}$ sous $H0$ $(p_{\min}^{(1)}, ..., p_{\min}^{(B)})$ et d'estimer la p-value globale résultante de l'approche :

$$p_G = \frac{\sharp\{p_{\min}^{(i)} \leqslant p_{\min}^{\text{obs}}\}}{B}.$$

Software

L'algorithme `LHiSA` est implémenté et disponible sous différentes versions[4] : la version en C++ a été réalisée de façon à s'exécuter rapidement tout en utilisant un minimum de mémoire. Pour cela, chaque génotype est codé dans 2 bits de mémoire, 0 et 1 correspondant aux allèles a et A respectivement ce qui permet de réduire considérablement l'espace mémoire utilisé. L'algorithme peut être lancé en ligne de commande ou à partir d'une application web réalisée par Mark Hoebeke. A titre indicatif, en procédant à un calcul des p-values par Monte-Carlo et sur une machine tout à fait raisonnable (Intel Pentium 4, CPU 2.80 GHz, 512 Mo RAM), l'implémentation en C++ prend à peu près 10, 180 et 800 secondes pour traiter 200, 2,000 et 10,000 SNPs respectivement. Par ailleurs une version en R (plus lente) existe et devrait bientôt être disponible de façon officielle.

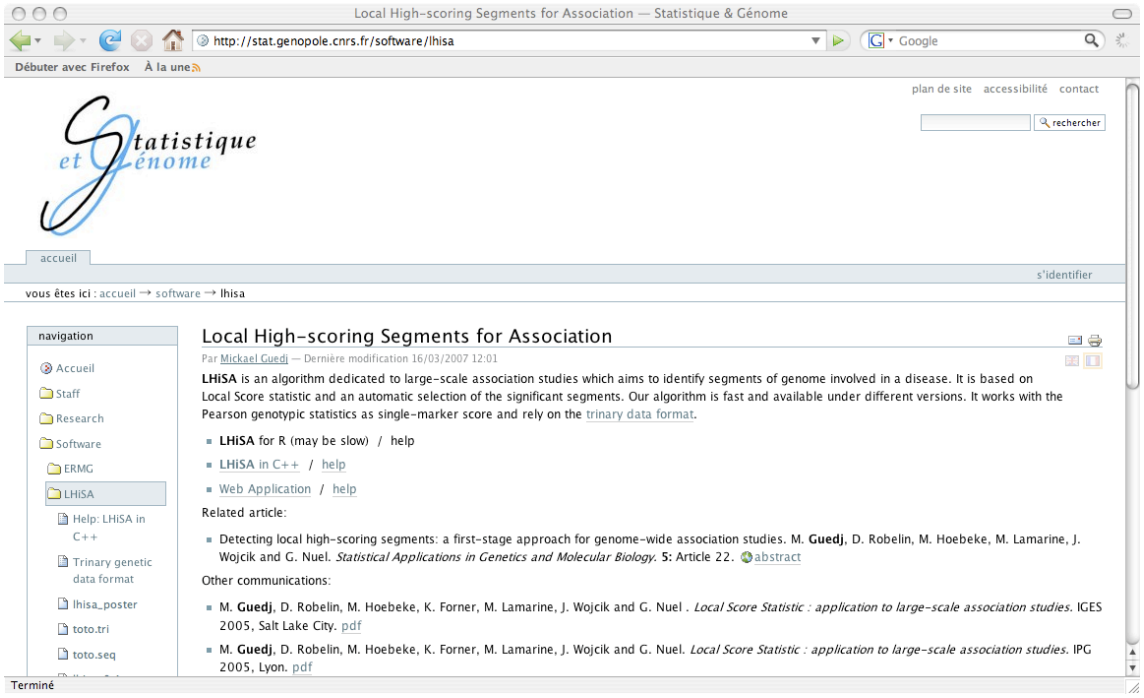
3.5 Applications

Application 1 : étude de puissance

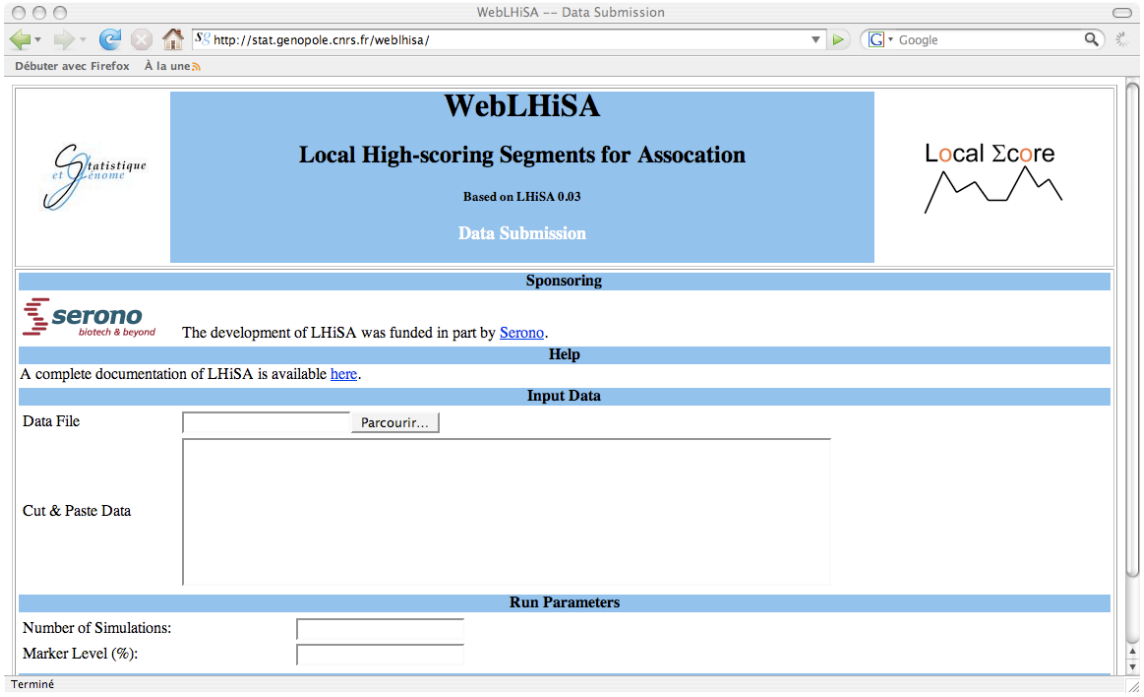
- Simulations : l'étude de puissance de cette nouvelle approche s'appuie sur des simulations de Monte-Carlo. La différence avec l'étude réalisée au chapitre 2 est qu'elle nécessite de générer un certains nombre de marqueurs consécutifs, liés les uns avec les autres par un *pattern* de LD le plus réaliste possible. Une façon de procéder consiste à appliquer, sur une distribution empirique des diplotypes possibles[5] obtenue à partir d'une population

[4] `http://stat.genopole.cnrs.fr/softwares/lhisa`

[5] ensemble de génotypes concernant deux locus consécutifs. Correspondant également à une paire d'haplotypes non-phasés.



http://stat.genopole.cnrs.fr/software/lhisa



http://stat.genopole.cnrs/weblhisa

réelle, le modèle génétique introduit au chapitre 2 page 57 afin de générer un jeu de cas et de témoins (Chapman et al 2003, Nielsen et al 2004). Ici nous prenons en compte les 674 SNPs du chromosome 19 génotypés pour une population française de 301 témoins en utilisant une puce *Affymetrix* 100K. Le locus de susceptibilité est choisi de façon équiprobable parmi les marqueurs pour lesquels la plus faible proportion génotypique excède 5%. Par ailleurs nous fixons la prévalence (K) à 0.05, la proportion allélique (p_A) à 0.3, le coefficient de consanguinité ($\mathcal{F}$) à 0 et le mode de transmission est additif. Le nombre de cas et de témoins est de 500 et le nombre de simulations de 200. Nous faisons varier la valeur du risque relatif RR_2 et du paramètre δ et nous avons considéré différentes situations pour lesquelles le locus de susceptibilité est contenu dans le jeu de données simulé ou non. Enfin, nous générons également des données pour 3 locus de susceptibilité indépendants, simulées à partir du même jeu de paramètres du modèle génétique. La statistique simple-marqueur utilisée est la statistique de Pearson appliquée sur la table des génotypes (X_G). Les résultats de puissance sont comparés à ceux obtenus en appliquant une approche simple-marqueur corrigée par Bonferonni ou Benjamini-Hochberg. Les détails sur ces deux corrections ont été donnés en introduction page 30.

- Résultats : les résultats sont présentés figure 3.8. Que le locus de susceptibilité soit inclus dans le jeu de données ou non, l'approche Score Local montre de meilleurs résultats que les analyses simple-marqueur corrigées par Bonferonni et Benjamini-Hochberg, affichant une augmentation moyenne de puissance de 0.25. Le fait de retirer le locus de susceptibilité des données simulées diminue naturellement la puissance respective de chaque approche. On peut aussi noter que la correction de Benjamini-Hochberg est plus puissante que celle de Bonferonni ce qui est en accord avec la littérature ; il faut cependant rappeler qu'elle implique également un taux d'erreur de type-I un peu plus important. Dans tous les cas simulés avec un locus de susceptibilité, le paramètre δ ne semble pas affecter la puissance de notre approche. De plus, sous $H0$ ($RR_2 = 1$), le taux d'erreur de type-I[6] est contrôlé à 5% quelle que soit la valeur de δ ce qui est une bonne façon de s'assurer de sa validité statistique.

Lorsque l'on traite les simulations réalisées avec trois locus de susceptibilité, le Score Local présente également de bonne performances comparés aux deux autres approches simple-marqueur. Dans ce cas, la valeur de δ ne semble pas avoir plus d'effet sur la probabilité de rejeter l'hypothèse nulle. En revanche, la capacité de la méthode à détecter des vrais découvertes estimée par la sensibilité c'est à dire le nombre de vrais-positifs sur le nombre de locus de susceptibilité ($= \frac{vp}{V}$ en reprenant les notations données en page 30) augmente avec δ ; le *True Discovery Rate* (TDR = 1−FDR) résultant estimé par le nombre de faux-positifs sur le nombre de positifs ($= \frac{fp}{R}$) diminue avec δ. Pour indication une région est considérée comme "vraie" si elle inclut un locus de susceptibilité et "fausse" sinon ; de la même manière elle est considérée comme "positive" si elle est inclue dans la sélection proposée par notre méthode et "négative" sinon.

[6]pour rappel il s'agit de la probabilité de rejeter à tord l'hypothèse $H0$: *il n'y a aucune région associée à la maladie*

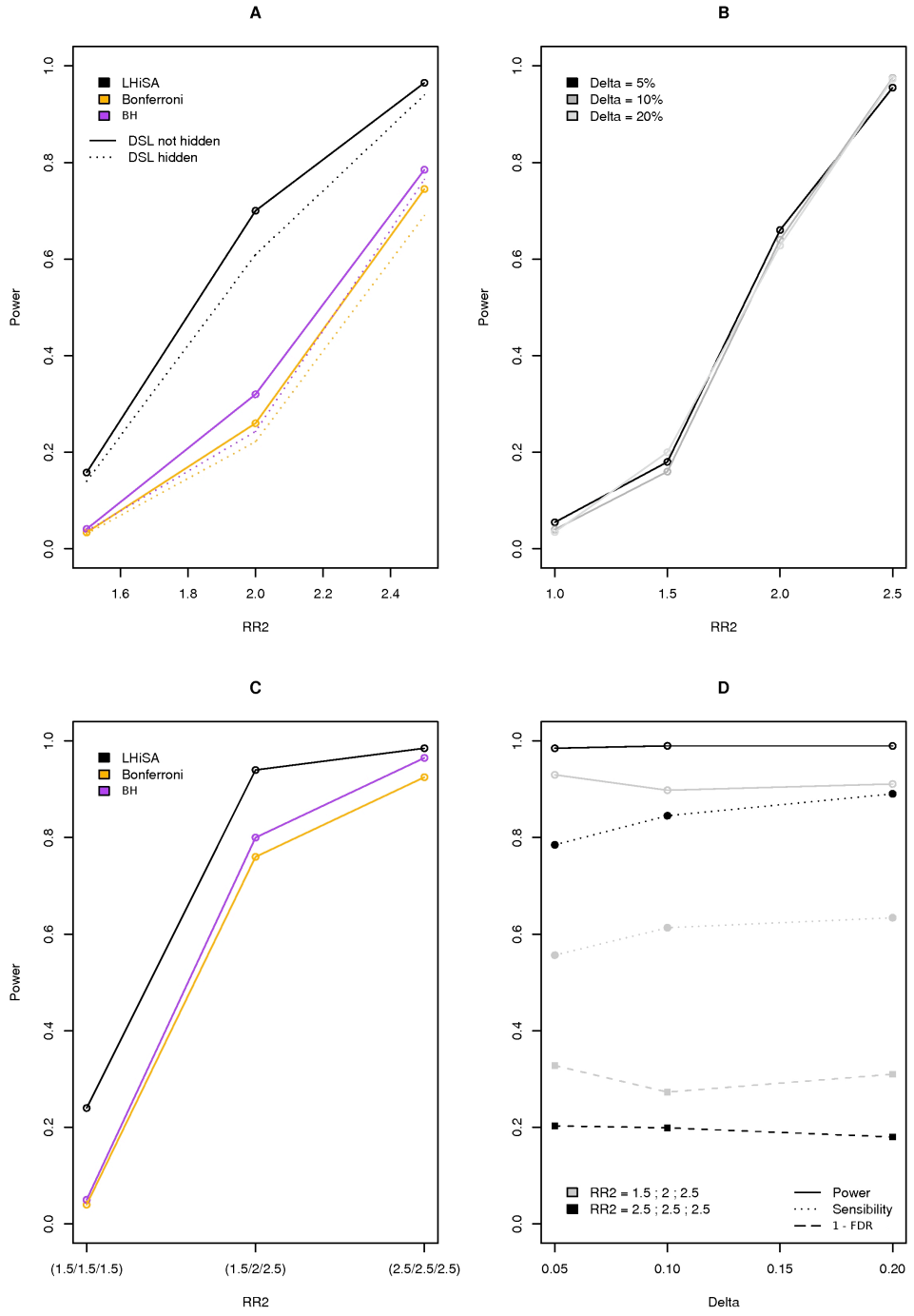


FIG. 3.8 – **Étude de puissance : A-** Comparaison entre LHiSA, Bonferroni et Benjamini-Hochberg pour un locus de susceptibilité et trois alternatives ($RR_2 = 1.5, 2$ et 2.5). **B-** Effet du paramètre δ sur la puissance de LHiSA pour un locus de susceptibilité, l'hypothèse nulle ($RR_2 = 0$) et trois alternatives ($RR_2 = 1.5, 2$ et 2.5). **C-** Comparaison entre LHiSA, Bonferroni et Benjamini-Hochberg pour trois locus de susceptibilité et trois alternatives ($RR_2 = \{1.5/1.5/1.5\}, \{1.5/2/2.5\}$ et $\{2.5/2.5/2.5\}$). **D-** Effet du paramètre δ sur la puissance, la sensibilité et le TDR (=1-FDR) résultant de LHiSA pour deux alternatives ($RR_2 = 1.5/2/2.5$ et $2.5/2.5/2.5$).

Application 2 : schizophrénie

- Données : nous appliquons notre approche sur des données d'association impliquant les gènes *G*72 et *DAAO*[7] dans la schizophrénie (Chumakov et al 2002). Le jeu de données consiste en 172 SNPs dont 8 sur le chromosome 12 et 164 sur le chromosome 13 dans des régions génomiques incluant *G*72 et *DAAO* respectivement. Les ADN de 213 patients et 241 témoins canadien ont été génotypés pour ce jeu de marqueurs. Pour l'approche Score Local, nous considérons la p-value du test allélique pour chaque SNP (pv_i) de sorte que $\mathcal{S}_i = -\log_{10}(\text{pv}_i)$ et $\mathcal{S}'_i = -\log_{10}(\text{pv}_i) - \delta$ avec $\delta = -\log_{10}(0.1)$ correspondant au seuil 10% pour la statistique d'association simple-marqueur considérée. Par ailleurs, l'implication de *G*72 dans la schizophrénie a été reportée à plusieurs reprises dans la littérature (Detera-Wadleigh et McMahon 2006).

- Résultats : l'approche Score Local identifie trois régions ($r = 3$) et la p-value correspondant à la somme des premiers Score Locaux vaut $p_T^{(3)} = p_{\min}^{\text{obs}} = 0.1660$. La significativité globale de l'approche p_G vaut 0.22. Les deux premières régions sont contenues dans *G*72 et la troisième est contenue dans *DAAO*. Notre méthode réussit donc à identifier les deux gènes attendus. Néanmoins, la significativité globale est loin d'être réellement convaincante quant à l'association réelle de cet ensemble de régions avec la maladie. Cela s'expliquer par l'effet modeste de ces deux gènes sur la maladie ou venir confirmer certains doutes émis récemment quant à l'implication de ces deux gènes dans la schizophrénie (Riley et Kendler 2006). Par ailleurs, même en plaçant le niveau individuel de chaque test à 0.22, les approches simple-marqueur corrigées par Bonferroni et Benjamini-Hochberg ne détectent aucun SNPs associés à la maladie, la plus faible p-value allélique observée par les auteurs étant de de 0.003 ($0.003 \times 172 = 0.516$).

rang	chr	région	H	T	p_T
1	13	149-153	2.542	2.542	0.2459
2	13	159-161	1.978	4.520	0.1737
3	12	5	1.165	5.686	0.1660
4	13	84	0.758	6.444	0.1702
5	13	49-51	0.587	7.031	0.1747

Tab. 3.1 – Résultats de LHiSA sur les données concernant la schizophrénie.

Application 3 : données AIM-Scan

- Données : nous appliquons notre stratégie aux données AIM-Scan *genome-wide* concernant la sclérose en plaque. Le jeu de données comprend 114,802 SNPs ; les ADN de 279 patients et 301 témoins suédois ont été génotypés en utilisant une puce *Affymetrix* 100K. La

[7] D-Amino Acid Oxidase

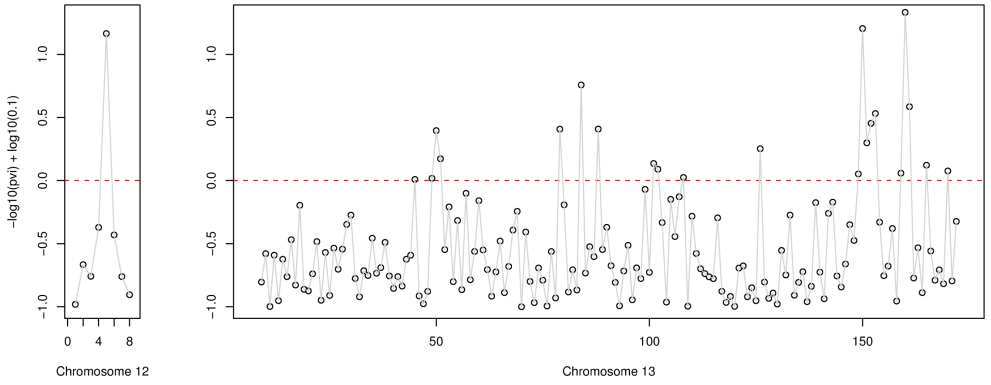


Fig. 3.9 – **Application sur la schizophrénie :** à chaque SNP est attribué une statistique d'association $\mathcal{S}_i = -\log_{10}(p\mathrm{v}_i)$ avec $p\mathrm{v}_i$ la p-value issue du test allélique. Les SNPs indéxés de 1 à 8 sont localisés sur le chromosome 12 et ceux indexés de 9 à 172, sur le chromosome 13.

statistique d'association associée à chaque marqueur s'appuie sur la p-value issue du test allélique non-biaisé et exact que nous avons développé (p. 69) : $\mathcal{S}_i = -\log_{10}(p\mathrm{v}_i)$. Le paramètre δ est fixé au niveau 5% pour la statistique d'association choisi : $\delta = -\log_{10}(0.05)$.

- Résultats : l'approche Score Local met en avant 3 régions recouvrant au total 70 SNPs. Parmi ces SNPs, 6 sont en commun avec ceux identifiés par l'analyse simple marqueur réalisée avec le même test et corrigée par Benjamini-Hochberg, inclus uniquement dans la première région. Par conséquent, le Score Local met en avant deux régions qui ne sont pas identifiées par l'analyse simple-marqueur. Réciproquement, l'analyse simple-marqueur implique 58 marqueurs dont la plupart ne sont pas identifiés par le Score Local. Par conséquent sur ce jeu de données, les deux approches donnent des résultats différents et éventuellement complémentaires.

Une des régions permet de localiser le gène *BTNL*2. Ce gène situé sur le chromosome 6 est plus connu pour son implication dans la maladie de Besnier-Boeck[8], une autre maladie auto-immune plus rare. Ce gène a été récemment mis en cause dans la sclérose en plaque mais on a montré que l'association observée avec la maladie était liée à un déséquilibre de liaison élevé entre ce gène et le locus *HLA-DRB*1 dont l'haplotype *HLA-DRB*1*15 est très fortement associé au risque de développer la sclérose en plaque (Traherne et al 2006).

[8] ou Sarcoidosis

Application 4 : données GAW15 et sévérité

- Données : à l'occasion de la 15e édition du *Genetic Analysis Workshop*, il a été distribué 100 réplications d'un jeu de données simulé, essayant de mimer des données concernant la polyarthrite rhumatoïde. Chaque réplication inclut 1,500 familles nucléaires et 2,000 témoins. Le jeu de marqueurs comprend 9,187 SNPs distribués le long de génome de façon à imiter une puce 10K. Ces données proposent par ailleurs la localisation des marqueurs ainsi qu'un certain nombre de covariables concernant les individus (sexe, âge ...). Parmi les 9 locus traits simulés, 2 appelés locus G et H localisés sur le chromosome 9 ont un impact important sur la sévérité à laquelle nous nous intéressons. La sévérité d'une maladie réfère à sa gravité et est indexée dans notre cas de 1 (peu sévère) à 5 (très sévère). Afin de détecter les marqueurs associés à la sévérité, nous considerons les individus affectés et extrêmes en terme de sévérité, c'est à dire ceux qui présentent un index de sévérité de 1/2 ou de 4/5. La statistique d'association ($\mathcal{S}_i$) utilisée ici est fondée sur la p-value (pv_i) issue du test allélique qui contraste cette fois-ci, non plus les cas des témoins, mais les individus présentant un indice de sévérité faible (1 ou 2) de ceux présentant un indice de sévérité élevé (4 et 5) : $\mathcal{S}_i = -\log_{10}(\text{pv}_i)$. Le paramètre δ est fixé au niveau 5% pour la statistique d'association choisie : $\delta = -\log_{10}(0.05)$.

Notre objectif ici est d'évaluer, sur 50 réplications, le nombre de fois que notre approche identifie les locus G et H et de comparer ses performances avec les approches simple-marqueur corrigées par Bonferroni et Benjamini-Hochberg. Les deux locus à identifier ayant été préalablement retirés des données, nous considérons que chaque approche détecte effectivement un locus si elle met en avant un des deux marqueurs flanquant ce locus.

- Résultats : on voit table 3.2 que l'approche Score Local détecte les deux locus dans la majeure partie des 50 réplications. Si les approches simple-marqueur détectent assez bien le locus G, les conclusions concernant le locus H ne sont pas aussi positives. L'avantage du Score Local réside ici dans la proximité des deux locus qui permet de renforcer la statistique, ce que ne prend pas du tout en compte une approche simple-marqueur.

locus	LHiSA	Bon	BH
G	49	42	44
H	45	5	8

Tab. 3.2 – **Performance sur les données GAW15 :** nombre de réplications sur les 50 réplications considérées où les locus G et H sont détectés par l'approche Score Local (LHiSA) et simple-marqueur corrigée par Bonferroni et Benjamini-Hochberg.

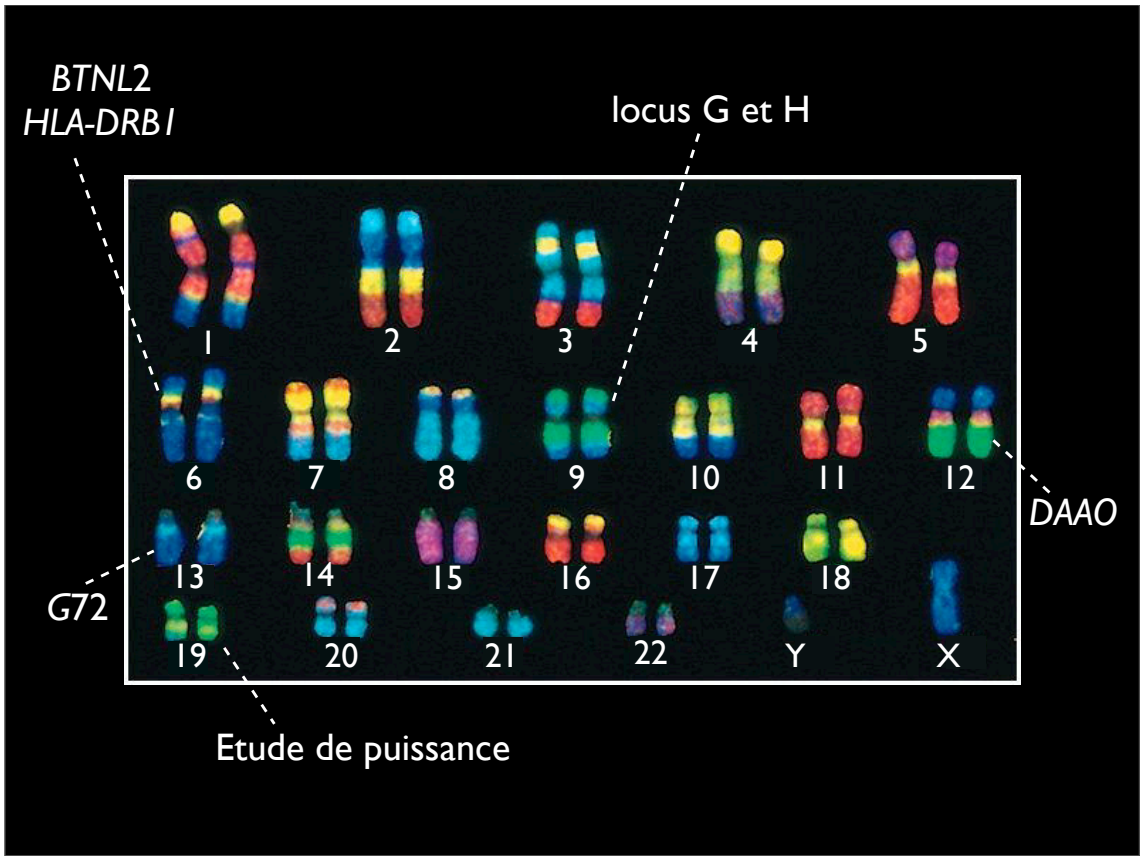


Fig. 3.10 – Locus et chromosomes impliqués dans les quatre applications du Score Local.

3.6 Discussion sur le Score Local

L'utilisation du Score Local dans les études d'association apparaît être une solution simple et rapide pour détecter des régions génomiques associées à la maladie, plutôt que des marqueurs individuellement. Reposant sur l'accumulation de hautes valeurs de statistiques d'association autour de un ou plusieurs locus de susceptibilité voisins, il permet de prendre en compte une dépendance locale qui s'exerce entre les marqueurs et d'identifier plusieurs régions distantes, voire sur des chromosomes différents. Cette approche peut donc être considérée comme multi-point et multi-locus. Elle peut également être apparentée aux approches par sommes de statistiques que nous avons introduites au début de ce chapitre (p. 107). L'avantage par rapport à l'approche Fenêtre Glissante est que le Score Local ne nécessite pas de contraindre à l'avance la taille des fenêtres ou régions. La principale différence avec la *Set Association* réside dans le fait que l'unité statistique de la méthode n'est plus le marqueur mais une région de marqueurs contigus. Sur un jeu de marqueurs indépendants et en l'absence d'agrégation de locus de susceptibilité dans une même région du génome, les deux approches devraient d'ailleurs fournir des résultats tout à fait comparables.

- Test-multiple : en réduisant le nombre de tests de n marqueurs à k régions, puis de k régions à un test pour l'ensemble de la procédure, l'algorithme que nous avons développé permet aussi de contourner le problème du test-multiple. On a également l'assurance que le taux d'erreur de type-I est contrôlé au niveau α que l'on se donne.

- Paramètre : notre méthode nécessite de fixer un seul paramètre : δ. Ce paramètre est facile à interpréter et donc à déterminer : il correspond au seuil simple-marqueur au dessus duquel on considère qu'un marqueur doit positivement contribuer au Score Local ; à l'inverse, un marqueur dont la statistique d'association est inférieure à δ le pénalise. L'effet de δ sur les résultats est assez naturel. Dans un premier temps la valeur de δ va affecter le nombre et la taille des régions : un δ élevé entraînera un nombre plus petit de régions et de tailles plus petites qu'un δ moins stringent. Nous avons également pu constater que si δ n'a heureusement aucun impact sur le taux d'erreur de type-I attaché à un niveau α donné, la puissance est également peu sensible au choix de ce paramètre. Il affecte néanmoins et assez logiquement la nature des résultats. En l'occurrence dans notre étude de puissance, nous constatons que le nombre de faux-positifs dans l'ensemble des régions proposées diminue avec δ tandis que le nombre de locus de susceptibilité effectivement localisés augmente.

- Flexibilité : enfin, un dernier avantage du Score Local est qu'il constitue un cadre d'analyse flexible qui peut être appliqué à n'importe quelle statistique d'association simple-marqueur ($\mathcal{S}_i$). Nous l'avons exploré dans notre application sur les données GAW en nous intéressant aux locus impliqués dans la sévérité de la maladie. On peut également imaginer

pouvoir détecter des marqueurs impliqués dans des phénomènes d'interactions gène-gène ou gène-environnement en construisant $\mathcal{S}_i$ sur la base du terme d'interaction issu d'une Régression Logistique par exemple, entre chaque marqueur et la covariable (génique ou environnementale) d'intérêt.

Implémentation : cependant, la flexibilité du Score Local constitue également un de ses principaux inconvénients. Il n'est en effet pas possible de proposer une implémentation universelle capable de répondre à tous les problèmes auxquels le généticien peut vouloir répondre, du fait d'un calcul de la significativité globale par Monte-Carlo qui dépend à la fois de la statistique $\mathcal{S}_i$ employée mais aussi de la structure même des données. De plus l'estimation des p-values correspondant aux sommes $T^{(i)}$ est également réalisée par Monte-Carlo plutôt qu'en utilisant les résultats issus de la théorie des valeurs extrêmes. Ceux-ci nécessitent en effet de faire l'hypothèse que les marqueurs sont indépendants ce qui n'est pas le cas en pratique et ce qui introduirait un biais dans le calcul des p-values. Les simulations de Monte-Carlo permettent d'intégrer le *pattern* de déséquilibre de liaison liant les marqueurs dans le calcul des p-values. Cela nous assure qu'une faible p-value est effectivement due à une association à la maladie, et non à une acumulation d'associations élevées observées par chance. En revanche, les simulations augmentent considérablement le temps d'exécution de la méthode. Cette dernière remarque est néanmoins limitée puisque comme nous l'avons vu, le temps total d'exécution reste raisonnable en pratique (3 minutes pour traiter 10,000 marqueurs par exemple).

- Sélection des régions : un problème sérieusement inquiétant quant à l'utilisation de notre approche réside dans le processus de sélection des régions. Elle repose sur la variation des p-values attribuées aux $T^{(i)}$. Dans la situation où les meilleures régions ont un effet extrêmement élevé sur la maladie[9], de telles variations ne sont pas décelables avec le nombre de simulations que l'on a l'habitude de réaliser (de l'ordre de quelques milliers). Cela a pour effet de sélectionner un nombre très important de régions avec une p-value estimée à 0. Le choix de notre méthode de sélection s'est inspiré de considérations non abouties formulées par Karlin et Altshul (1993). Il est cependant possible de considérer d'autres méthodes qui sélectionnent les régions en s'appuyant sur d'autres critères que la variation des p-values attribués aux $T^{(i)}$ tels que la distribution du nombre de Scores Locaux par exemple. Ce point est développé dans le dernier chapitre du manuscrit.

- *Pattern* de LD : l'utilisation du Score Local nécessite par ailleurs que l'intensité du déséquilibre de liaison soit relativement homogène entre les marqueurs le long du génome. Dans le cas contraire, la détection d'un locus de susceptibilité en déséquilibre de liaison avec beaucoup de marqueurs est favorisée par rapport à celle d'un locus de susceptibilité en déséquilibre de liaison avec peu de marqueurs. Cela étant, les approches simple-marqueur reposent également sur la qualité du déséquilibre de liaison entre les locus de susceptibilité

[9] comme cela peut-être le cas avec la région HLA du chromosome 6 pour les maladies auto-immunes

et les marqueurs génotypés *via* le principe d'association indirecte. De plus, l'approche par Score Local permet ainsi de mettre en compétition des locus de susceptibilité isolés qui ont un effet majeur sur la maladie avec ceux qui ont un effet plus modeste mais qui seraient "mieux" entourés par le jeu de marqueurs ou qui s'agrègeraient dans une région génomique donnée, ce que ne permet pas une approche simple-marqueur.

Interaction : enfin il est important de noter qu'à l'instar des approches multi-marqueurs qui s'appuient sur des sommes de statistiques, l'approche par Score Local n'est absolument pas dédiée à la détection d'effets d'interaction entre les marqueurs ce qui peut être vu comme un inconvénient, et discuté dans la dernière section de ce chapitre.

Application : si les simulations et quelques unes de nos applications ont permis de mettre en avant l'intérêt du Score Local par rapport aux approches simple-marqueur, l'attitude à adopter en pratique nécessite néanmoins plus d'expérience. Il se peut dans certaines situations, que le Score Local n'apporte pas plus d'information que les approches simple-marqueur. Dans d'autres, il peut aboutir à des conclusions totalement différentes. Dans un tel cas, on peut considérer que les différentes approches apportent des résultats complémentaires. Ce dont nous pouvons être sûrs, c'est que les différences de résultats seront essentiellement dues au *pattern* de LD entre les locus de susceptibilité et le jeu de marqueurs génotypés, ainsi qu'à l'éventuelle proximité des locus de susceptibilité.

3.7 Conclusions

Ce chapitre a été l'occasion de proposer une *review* des principales approches multi-locus publiées dans la littérature et de les discuter. Une partie importante du travail réalisé s'est également concentrée sur le développement d'une nouvelle méthode reposant sur la statistique du Score Local, initialement utilisée dans l'analyse des séquences biologiques. Dans le contexte des études d'association *genome-wide*, l'objectif est de proposer un outil simple et rapide pour détecter des régions génomiques associées à la maladie.

Si le travail réalisé sur cette méthode est conséquent en terme de développement et d'application, il reste néanmoins un certain nombre de points à approfondir sérieusement. En l'occurrence, il est impératif de l'appliquer plus largement sur des études d'association à grande échelle de façon à affiner la compréhension et l'interprétation des différences observées entre les résultats obtenus avec cette méthode, et ceux obtenus avec une approche simple-marqueur. Il nous paraît également important de réaliser plus de simulations afin d'évaluer le Score Local dans des situations qui impliquent un plus grand nombre de locus de susceptibilité ainsi que des *patterns* de LD différents. Il peut par ailleurs être intéressant d'intégrer dans la délimitation des régions des informations d'annotation telles que les limites des gènes par exemple ; à ce sujet, on peut noter que le logiciel que nous proposons

prend d'ores et déjà en compte les limites entre les chromosomes, une région à cheval entre deux chromosomes ne correspondant naturellement à aucune réalité biologique. Enfin, le processus de sélection que nous avons choisi à l'étape 3 de l'algorithme n'est pas sans soulever des problèmes. Ce point technique est sans doute pour nous le plus important à approfondir. Il est discuté avec plus de détails dans le dernier chapitre de ce manuscrit.

Chacune des méthodes multi-locus que nous avons évoquées présente ses avantages et ses inconvénients. Celle que nous avons construite à partir du Score Local n'échappe pas à cette constatation. Afin de la replacer dans le contexte actuel, nous l'avons ajoutée au tableau d'évaluation que nous avons proposé en début de chapitre (figure 3.11 p. 136). De façon très générale, on peut conclure que chacune des méthodes multi-locus exprime ses performances par rapport à un compromis entre la possibilité de traiter un grand nombre de marqueurs et celle de considérer des combinaisons entre marqueurs d'ordre élevé. La statistique du Score Local par exemple n'est pas adaptée à la détection d'interactions. En contrepartie, la MDR a été développée dans le but de détecter des combinaisons de génotypes associés à la maladie mais ne peut pas gérer un jeu de données de l'ordre de la centaine de milliers de marqueurs en termes de temps d'exécution. Une idée peut donc être d'allier les performances des deux méthodes en identifiant dans un premier temps les régions intéressantes avec le Score Local et en appliquant dans un deuxième temps la MDR sur les marqueurs localisés dans ces régions.

	RL	FG	SA	MDR	FA	LHiSA
Dimensions	+	+++	+++	+	++	+++
LD	non	oui mais ...	non	non	non	oui
Simplicité	++	+++	+++	+	+	+++
Choix des paramètres	++	++	++	+	+	+++
Résultats	+++	+++	+++	+	+	+++
Interactions	oui	non	non	oui	oui	non
Effets modestes	oui	non	non	oui	oui	non mais ...

Fig. 3.11 – **Critères d'évaluation :** compare les performances de cinq méthodes multi-marqueurs sur la base de sept critères d'évaluation. Les méthodes sont la Régression Logistique (LR), la Fenêtre Glissante (FG), la *Set Association* (SA), la *Multiple Dimensionary Reduction* (MDR), la Forêt Aléatoire (FA) et la méthode reposant sur le Score Local (LHiSA). Par ailleurs (+), (++) et (+++) représentent une note moyenne, bonne et très bonne respectivement.

Chapitre 4

Conclusions

Avec l'avancement des connaissances en Biologie Moléculaire, les études d'association à grande échelle sont aujourd'hui devenues incontournables dans le contexte de L'Épidémiologie Génétique moderne. Elles offrent la possibilité de s'intéresser simultanément à des centaines de milliers de variants génétiques dans une population donnée, afin d'identifier les quelques variants impliqués dans l'apparition de maladies. Devant ce problème qui peut être imagé par *"chercher quelques aiguilles dans une botte de foin"*, l'apport de méthodologies nouvelles et adaptées prend une part de plus en plus importante dans la littérature associée à ce type d'études, comme cela s'est produit précédemment pour l'analyse des données issues de la génomique et de la post-génomique. Cette thèse s'inscrit dans ce contexte : l'objectif à l'origine est de répondre aux problématiques statistiques soulevées par l'industriel *Serono* dans un domaine de recherche nouveau pour le laboratoire *Statistique et Génome* : L'Épidémiologie Génétique. Ce travail s'est donc d'abord traduit par une recherche bibliographique conséquente, permettant de pointer l'ensemble des thématiques associées aux études d'association *genome-wide*. Parallèlement, nous nous sommes efforcés de proposer un développement méthodologique pertinent, supporté par la mise à disposition au sein de la communauté des algorithmes correspondants.

Ce chapitre porte naturellement sur les conclusions de cette thèse en reprenant l'ensemble des aspects importants mais il est également l'occasion d'évoquer les différentes perspectives scientifiques qu'elle ouvre. Enfin nous clôturons ce manuscrit par une discussion générale des études d'association *genome-wide* ainsi qu'une ouverture sur les bénéfices certains apportés par l'intégration en Épidémiologie Génétique de thématiques liées à l'étude des données génomiques et post-génomiques.

4.1 Conclusions générales

L'une des ambitions de ce manuscrit est de constituer une lecture de référence pour quiconque voudrait faire un premier pas dans l'analyse des études d'association *genome-wide*. L'introduction en rend particulièrement compte. En plus de rappeler les préceptes statistiques et génétiques fondamentaux et nécessaires à la compréhension de cette thèse[1], nous avons abordé L'Épidémiologie Génétique en évoquant et discutant les différents *designs* d'étude possibles[2], ainsi que leur déroulement, de la génération des données à la formulation et la confirmation d'hypothèses. En particulier, nous avons voulu souligner les problèmes que soulève le principe de réplication, inextricablement liés à la qualité des données et des résultats. Le manque de puissance, le test-multiple, les erreurs de génotypages, les valeurs manquantes et la stratification éventuelle de la population sont autant de facteurs à prendre en considération au moment de l'analyse de façon à pouvoir rendre compte de la validité et de la fiabilité des résultats obtenus.

Le corps de ce manuscrit tourne autour des deux grands axes de l'analyse statistique de données d'association : l'analyse simple-marqueur et l'analyse multi-marqueurs.

L'analyse simple-marqueur vise à identifier les marqueurs associés à la maladie en leur appliquant individuellement le même test d'association. Bien que l'efficacité d'une telle stratégie puisse être remise en question dans le contexte des maladies complexes, l'analyse simple-marqueur reste en pratique l'approche privilégiée. A première vue tout à fait triviale, elle soulève néanmoins un certain nombre de problématiques. Au regard du nombre important de marqueurs à tester et des effets modestes que l'on cherche à détecter, il est important que l'analyse simple-marqueur soit menée avec précaution quant aux possibles biais qui peuvent affecter la qualité des résultats. Outre des considérations d'ordre statistique tels que la nature du test utilisé pour tester la même hypothèse nulle[3], le mode d'estimation de la p-value[4] et le modèle d'échantillonnage utilisé[5], nous avons vu qu'il existait différents tests construits sur des hypothèses nulles et alternatives sensiblement différentes. Les tests génotypique et allélique cherchent à mettre en évidence une différence entre les proportions génotypiques ou alléliques observées et celles attendues sous l'hypothèse d'indépendance entre un marqueur et la maladie. Le test de tendance prend en compte un ordre dans les génotypes et suppose que la probabilité d'être affecté augmente linéairement avec le nombre d'allèles de susceptibilité (modèle additif). Enfin, le test d'Hardy-Weinberg s'appuie sur la déviation par rapport à l'équilibre observée chez les cas. Devant cette liste non exhaustive de tests possibles, on peut se demander quelle stratégie il convient de mettre en place. Pour cela, nous avons réalisé une étude de puissance ; à cette occasion nous avons défini un modèle génétique de façon obtenir des alternatives réalistes, et nous avons discuté différents cadres statistiques permettant d'approcher la

[1] tels que le test d'hypothèse, le déséquilibre de liaison et l'équilibre d'Hardy-Weinberg
[2] familiale/cas-témoins, liaison/association, gènes-candidats/*genome-wide*
[3] test de score, test de vraisemblance ou test de Wald
[4] asymptotique, empirique ou exact
[5] test conditionnel, non-conditionnel

distribution des statistiques considérées sous ces alternatives. A travers notre étude de puissance, nous avons vu que l'efficacité de chaque test dépend naturellement du modèle génétique sous-jacent. Ceux-ci n'étant en pratique pas connus, une stratégie séduisante consiste alors à combiner ces tests ou les statistiques correspondantes dans le but d'obtenir la meilleure puissance en toute circonstance. On parle alors de méta-statistiques. Si nous avons observé que cette stratégie n'est la meilleure dans aucune des alternatives considérées, elle permet en revanche d'atteindre un niveau de puissance intermédiaire et souvent plutôt proche du test le plus puissant. Mais du fait de la dépendance entre les différentes statistiques d'association, construire un test d'association sur la base d'une combinaison de tests ou de statistiques nécessite d'estimer la p-value avec précaution : on pourra obtenir la distribution sous l'hypothèse nulle de ces méta-statistiques par Monte-Carlo ou également, dans le cas de combinaisons linéaires de statistiques, par la Forme Quadratique introduite lors de l'estimation de la puissance.

Par ailleurs, un thème en filigrane tout au long de la première moitié du manuscrit traite de l'utilisation et l'interprétation des déviations observées par rapport à l'équilibre d'Hardy-Weinberg dans les études d'association. Nous avons vu en introduction qu'une déviation chez les témoins peut souligner des erreurs de génotypage ce qui fait du test d'Hardy-Weinberg le moyen le plus simple de traiter ce problème. Dans le chapitre 2, nous avons constaté qu'une déviation chez la population combinée des cas et des témoins peut remettre complètement en cause la validité du test allélique et qu'une déviation chez les cas uniquement permet de détecter de l'association et d'augmenter la puissance des autres statistiques d'association, conditionnellement au fait que la population générale est à l'équilibre. Par conséquent l'analyse de l'équilibre d'Hardy-Weinberg dans un jeu de données d'association peut à la fois renforcer l'analyse simple-marqueur tout comme la remettre en cause. Les études d'association pourraient tirer avantageusement parti d'un rapport plus systématique des ces informations en termes de qualité des résultats. Concernant la validité du test allélique, une façon simple de prendre en compte une déviation par rapport à l'équilibre d'Hardy-Weinberg est d'utiliser un test qui ne fait pas d'hypothèse sur l'appariement indépendant entre les allèles tels que le test de tendance ou le test allélique non-biaisé que nous avons proposé sous une déclinaison de test exact.

Au regard des relations qu'entretiennent les marqueurs et les locus de susceptibilité, à savoir principalement l'association allélique résultant du déséquilibre de liaison ainsi que d'éventuels effets d'interaction, les approches simple-marqueur apparaissent limitées pour prendre en compte toute la complexité sous-jacente aux jeux de données. Les statisticiens et les informaticiens se sont donc penchés sur des approches multi-marqueurs qui ne sont cependant pas sans poser des problèmes méthodologiques évident liés à la dimension des données et au niveau de complexité que l'on veut pouvoir traiter. Nous avons distingué deux familles d'approches : les approches multi-points qui considèrent plusieurs marqueurs contigus dans le but d'améliorer la détection ou la localisation d'un locus de susceptibilité, et les approches multi-locus qui considèrent un ensemble de marqueurs dans le but d'identifier plusieurs locus de susceptibilité, potentiellement distants. C'est à ces dernières que nous nous sommes intéressés. Les approches par régression logistique, par sommes de statistiques, combinatoires et par partitionnements récursifs que

nous avons évoquées constituent les principales solutions proposées dans la littérature. Nous proposons une nouvelle méthode qui s'appuie sur la statistiques du Score Local assimilée aux approches par sommes de statistiques. Elle a été développée à partir de la volonté d'effectuer sur une centaine de milliers de marqueurs ce que l'oeil réalise intuitivement sur une centaine : identifier des accumulations de valeurs élevées dans le signal d'association. De telles accumulations peuvent être dues au déséquilibre de liaison ou à une agrégation locale des locus de susceptibilité. Procédant ainsi, le Score Local permet d'élever l'unité de l'analyse du marqueur à toute une région, dont la taille n'est pas fixée à l'avance. Il sélectionne au final un ensemble de régions significativement associées à la maladie, qui pourront être par la suite étudiées avec plus de précision. Comparé aux autres approches, il est clair que chacune possède ses avantages et inconvénients, favorisant en général la possibilité de traiter un grand nombre de données ou un degré de complexité élevé. Il n'est pas vraiment possible de comparer leur performances respectives. Le Score Local par exemple n'a pas pour objectif d'identifier des interactions entre les locus. On peut néanmoins penser que les régions identifiées par cette méthode impliquent des locus de susceptibilité interagissant les uns avec les autres qui pourraient être identifiés en appliquant la MDR à la suite du Score Local.

Les deux principaux axes de ce manuscrit, à savoir les approches simple-marqueur et multi-marqueurs peuvent paraître au premier abord déconnectées. En réalité, les approches multi-marqueurs dépendent énormément de la stratégie adoptée en simple-marqueur. Déjà parce qu'une partie des méthodes considérant plusieurs marqueurs simultanément ne peuvent pas gérer le nombre important de marqueurs à traiter. L'analyse simple-marqueur peut alors être vue comme une pré-sélection des marqueurs à considérer par la suite. Ensuite parce que certaines méthodes telles que le Score Local reposent directement sur le test d'association choisi en simple-marqueur. La pertinence et la validité de l'analyse simple-marqueur a donc toutes les chances d'influencer par la même occasion l'efficacité des approches multi-marqueurs.

Quelle que soit la stratégie retenue, l'analyse statistique des données d'association nécessite d'effectuer un nombre conséquent de tests. Décider du seuil de rejet de l'hypothèse nulle à partir du contrôle du taux d'erreur de type-I associé à chaque test individuellement n'est pas adapté et entraîne un nombre trop important de faux-positifs lorsqu'ils sont contrôlés au niveau traditionnel de 5%. Il s'agit du problème du test-multiple. Une réponse à ce problème consiste à contrôler d'autres quantités statistiques relatives à la famille de tests réalisés. Le FWER contrôle la probabilité d'obtenir au moins un faux-positif. Le FDR contrôle la proportion de faux-positifs dans l'ensemble des positifs. L'utilisation de l'un ou de l'autre dépend en fait de la question posée. Le contrôle du FWER est souvent considéré comme trop conservatif et donc peu puissant pour identifier les locus de susceptibilité. Par le contrôle du FDR, on accepte une proportion de faux-positifs un peu plus importante en échange d'une meilleure sensibilité à détecter les vrais-positifs. Mais si par exemple, dans une optique de contrôle qualité, on décide de retirer de l'étude les marqueurs qui s'écartent significativement de l'équilibre d'Hardy-Weinberg chez les témoins, l'objectif est d'en retirer le moins possible. Le contrôle du FWER apparaît alors tout à fait adapté. Une troisième quantité a été introduite récemment et correspond à la probabilité

pour chaque marqueur d'être ou non associé à la maladie. Il s'agit du FDR Local. Si son estimation s'avère moins évidente à première vue que les majorations du FWER et du FDR proposées par Bonferroni et Benjamini-Hochberg respectivement, elle n'en reste pas moins, dans le cadre du modèle de mélange gaussien que nous avons présenté, simple et rapide. Dans le contexte actuel, le FDR Local apporte une information supplémentaire aux résultats, plus intuitive à interpréter pour le généticien que la p-value, le FWER ou même le FDR. Une manière alternative de répondre au problème du test-multiple consiste à réduire le nombre de tests en s'appuyant sur des informations biologiques pertinentes. L'existence le long du génome de blocs de LD en est une : en prenant en compte l'accumulation résultante de statistiques d'association élevées, le Score Local permet dans un premier temps de réduire le nombre de tests de n marqueurs à k régions, puis de k régions à un test, contournant ainsi le problème lié au test-multiple. D'autres stratégies suivent la même idée.

Les problématiques soulevées dans cette thèse ont été principalement guidées par les données d'association *genome-wide* cas-témoins proposées par Serono. La plupart sont néanmoins également valables dans le cadre d'études d'association familiales. Les différents modes d'estimation de puissance explorés peuvent s'appliquer également aux tests d'association construits sur des familles tels que le TDT, tout comme les conclusions sur les méta-statistiques : il y a par exemple fort à penser que la combinaison du test d'Hardy-Weinberg avec le TDT sur des données familiales puisse augmenter la puissance, tout comme cela se produit avec le test de tendance sur des données cas-témoins. Enfin le FDR Local et le Score Local impliquent des méthodologies qui peuvent s'appliquer à n'importe quelle statistique d'association, et donc également à celles issues d'études d'association familiales.

Le travail réalisé a également été supporté par le développement de logiciels : `fueatest` en ce qui concerne le test allélique exact non-biaisé, `kerfdr` pour l'estimation semi-paramétrique du FDR Local (en développement), et `LHiSA` pour l'algorithme construit à partir du Score Local. Tous ces logiciels sont développés dans le souci constant de répondre aux contraintes qu'imposent le traitement de données aussi importante en termes de temps d'exécution et de ressource mémoire utilisée. Les nombreuses considérations algorithmiques concernant `fueatest` et `LHiSA` en témoignent. Par ailleurs, il nous a semblé important de proposer chaque implémentation sous différentes versions (C++, R, Perl, Application Web) dans le but de faciliter leur application.

Enfin, nous avons abordé les problèmes éthiques liés au génotypage à grande échelle de populations humaines. Bien que ce point ait été traité succinctement en introduction, il nous paraissait important d'en parler dans le cadre d'un travail effectué à partir de telles données.

4.2 Perspectives

Le travail de recherche décrit dans ce manuscrit ne prend naturellement pas fin avec l'écriture de celui-ci. Il constitue pour nous une ouverture sur de nouvelles thématiques. Nous présentons ici la direction scientifique que nous prenons actuellement. Les points que nous voulons développer concernent essentiellement la poursuite des travaux réalisés sur le Score Local et sur le FDR Local ainsi que la prise en compte de l'hétérogénéité intrinsèque aux données d'association.

Score Local

Sur notre méthode, le point technique qui demande encore le plus de considération réside dans la sélection des régions retenues au final. Inspirée par de précédentes remarques formulées par Karlin et Altshul, cette sélection s'appuie sur la variation de la significativité des sommes de Scores Locaux. Nous avons vu que pour des régions très associées, cette manière de procéder a comme inconvénient de nécessiter un nombre très important de simulations, augmentant sensiblement le temps d'exécution. Nous proposons deux pistes à explorer pour contourner ce problème. La première consiste à tester séquentiellement les régions l'une après l'autre avec un argument de type *on sélectionne les régions tant qu'elles sont significatives à un niveau fixé*, comme cela est proposé dans la thèse de David Robelin. Cette démarche de tests séquentiels sur des variables ordonnées (ici les Scores Locaux successifs) a l'avantage de contrôler simplement le taux d'erreur de type-I au niveau fixé. La deuxième piste consiste à sélectionner les régions intéressantes sur la base de la distribution du nombre de régions sous l'hypothèse nulle. En effet, les différentes applications du Score Local nous ont permis de constater que le nombre de Score Locaux obtenus pour une valeur de δ donnée sous l'hypothèse alternative, était supérieur au nombre attendu sous l'hypothèse nulle. Par conséquent la différence entre les deux peut potentiellement nous fournir une estimation du nombre de régions à sélectionner. Notre objectif est donc dans un premier temps d'obtenir la distribution du nombre de Score Locaux, puis de comparer les trois approches en termes de puissance, de sensibilité et de spécificité des résultats. Notons que cette problématique a des applications plus larges que les études d'association puisque la question de la sélection des régions associées aux meilleurs Scores Locaux reste également ouverte dans le contexte de l'analyse des séquences biologiques.

Un autre point à développer concerne le post-traitement des régions sélectionnées par le Score Local. L'orientation proposée avec Sophie Lèbre est de traiter les marqueurs inclus dans les régions à l'aide d'un modèle graphique de façon à identifier dans les régions, les paquets de marqueurs associés au même locus de susceptibilité (blocs de LD) et entre deux régions, les locus de susceptibilité présentant des effets d'interaction.

Enfin, dans le contexte des études d'association *genome-wide* une thèmatique de re-

cherche consiste à déterminer le *design* d'étude optimal permettant de réduire le nombre de génotypages à réaliser tout en maintenant un niveau de puissance raisonnable. Dans cette optique, un grand nombre de *designs* ont été proposés, impliquant généralement une analyse en deux étapes : une étape pré-sélectionne à partir d'un nombre réduit d'idividus les marqueurs "intéressants" et une deuxième analyse sur un plus grand nombre d'individus, ce jeu réduit de marqueurs. En collaboration avec Hugues Aschard, nous travaillons actuellement sur une telle approche où la sélection à l'étape 1 est réalisée par le Score Local. Cette approche a été retenue à l'occasion du *Genetic Analysis Workshop* (GWA15) pour la session "*Novel methods*" et publiée dans *BCM Genetics* (Aschard et al 2007).

FDR Local

Ce que nous avons présenté sur le FDR Local représente pour nous un travail introductif sur cette thématique. Notre objectif actuellement est d'achever avant tout le développement de `kerfdr` en collaboration avec l'équipe de Stéphane Robin. Par ailleurs, nous avons vu que la transformation probit pouvait poser un problème de troncature pour les valeurs de p-values à 1 comme pour les p-values estimées à 0 résultant d'une estimation par Monte-Carlo. Nous considérons actuellement ce problème et comptons intégrer sa solution dans `kerfdr`. Enfin, la méthode d'estimation semi-paramétrique du FDR Local nécessite l'estimation *a priori* des proportions de marqueurs sous l'hypothèse nulle et sous l'hypothèse alternative ce que permet le modèle de mélange gaussien que nous avons présenté. Nous projetons d'en évaluer la qualité et de la confronter aux autres méthodes d'estimation proposées dans la littérature.

Si l'estimation du FDR Local peut permettre celle du FDR, il n'exclut par ailleurs pas de procéder à des estimations plus directs et éventuellement plus précises de ce dernier. L'équipe de Bioinformatique de *Serono* s'est penché sur une méthodologie d'estimation du FDR fondée sur la distribution empirique des statistiques d'association et adaptée à tout type de *design* d'étude. Ce travail est accepté dans *Human Heredity*.

Hétérogénéité intra-population : modèles LC-MELR

En introduction de ce manuscrit, nous avons vu que plusieurs facteurs tels que l'hétérogénéité allélique ou de locus pouvaient engendrer une population de cas inhomogènes en termes d'étiologie et conduire à une diminution sensible de la puissance.

Une approche bien établie pour prendre en compte une telle structure entre les individus est d'introduire dans le modèle des effets aléatoires en plus des effets dits fixes. Les modèles de Régression Logistique à Effets Mixtes (MELR) peuvent être utilisés pour prédire des variables discrètes dans des observations corrélées suivant une structure de corrélation donnée. Une extension intéressante des modèles MELR sont les modèles MELR

à Classes Latentes (LC-MELR) pour lesquels la structure de corrélation n'est pas connue et est inférée parallèlement à l'estimation des paramètres du modèle.

De tels modèles sont couramment utilisés en Sciences du Comportement mais apparaissent également attractifs dans les études génétiques où un grand nombre de variables ne sont pas observées. Nous avons parlé de l'hétérogénéité des étiologies, mais il peut également s'agir de variables environnementales ou génétiques telles que les haplotypes.

Hétérogénéité inter-populations : Réplications Locales

Dans les études génétiques, nous avons vu que la réplication des résultats dans une ou plusieurs populations indépendantes, était considérée comme l'approche privilégiée pour distinguer les faux-positifs des vrais signaux d'association. En pratique, de telles réplications s'observent difficilement. Au delà des facteurs liés à la qualité des données et des résultats, des conclusions contradictoires entre deux populations peuvent provenir de réelles différences biologiques entre les populations. En particulier, la variation du *pattern* de déséquilibre de liaison et des proportions alléliques entre des populations d'origines différentes pose un défi majeur pour la découverte de nouveaux locus de susceptibilité.

Plutôt que de s'intéresser à des réplications au sens strict du terme (même locus, mêmes allèles associés), nous proposons de considérer la réplication de régions plutôt que des marqueurs individuellement. On introduit à cette occasion le concept de Réplication Locale, définie par l'accumulation dans une région chromosomique localisée de hautes valeurs d'association, et répliquée entre les populations, sans forcément contraindre la réplication de tel ou tel marqueur. La détection de telles Réplication Locales peut se traduire un nouvelle fois en termes de recherche de Score Locaux. L'approche a été présentée lors du dernier *European Mathematical Genetics Meeting* (2007) et les premiers résultats nous confortent dans l'idée de persévérer dans cette direction.

4.3 De l'Épidémiologie Génétique à l'Épidémiologie Génomique

"Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases ?" Bourgain et al (2006).

Les études d'association *genome-wide* sont aujourd'hui devenues une réalité et se sont imposées pour la découverte de déterminants génétiques impliqués dans des maladies complexes. Les premiers résultats ont été publiés récemment (Maraganore et al 2005, Herbert et al 2006). Elles soulèvent cependant un grand nombre de problématiques liées au *design* de l'étude, à la qualité des données, des résultats et à la complexité des étiologies

auxquelles on s'intéresse. Chacun de ces points a été discuté au cours de ce manuscrit. Par ailleurs, les études d'association *genome-wide* n'excluent pas les études de liaisons ou gènes-candidats, cette diversité de stratégies répondant positivement à la diversité avec laquelle les locus de susceptibilité impactent sur les maladies.

L'objectif de telles études est de contribuer significativement à la connaissance des mécanismes biologiques à l'origine des maladies. Une fois un signal d'association obtenu avec suffisamment de certitude, le passage de cette association statistique à la compréhension des effets causaux impliqués est cependant loin d'être évident. Un exemple est donné par l'association de la région HLA dans le chromosome 6 avec les maladies auto-immunes. Si l'implication d'HLA est suggérée depuis plus de 20 ans, ses implications fonctionnelles dans des processus pathologiques restent encore peu connues.

Cet exemple soulève la nécessité de dépasser le niveau de recherche considéré par la Génétique - la variation de la séquence d'ADN entre les individus - afin d'atteindre le degré de complexité des organismes vivant en intégrant d'autres types de données issues de niveaux biologiques différents. On sait que l'ensemble de l'information contenue dans le génome ne se résume pas uniquement aux gènes mais progresse de l'ADN à la protéine en passant par un niveau intermédiaire : l'ARN[6]. Les avancées récentes en Génomique (relatif à l'ADN), Transcriptomique (relatif à l'ARN) et Protéomique (relatif aux protéines) ont permis, parallèlement à l'accumulation de données génétiques, de générer un grand nombre de données génomiques et post-génomiques indispensables à la compréhension des mécanismes biologiques à l'origine des maladies. Chaque niveau de recherche apportant une information différente ou partiellement redondante, les chercheurs peuvent tirer avantageusement parti de les considérer simultanément.

Identifier les zones génomiques associées à une maladie et similaires entre différentes espèces par Génomique Comparative dans le but de mettre en évidence des éléments de régulations impliqués, identifier les variants génétiques associés à l'expression différentielle de certains gènes et mettre en relation les réseaux d'interactions - au sens statistique - entre gènes avec les réseaux d'interactions - au sens fonctionnel - entre protéines, sont autant d'exemples qui montrent à la fois la diversité des possibilités d'analyse offertes aux chercheurs, mais aussi le défi que représente l'intégration de données biologiques hétérogènes (Reif et al 2004, Ibrahim et Gold 2005, Xiong et al 2005, Parsons et al 2005).

Dans ce contexte, le développement de méthodologies statistiques et informatiques adaptées est une nouvelle fois nécessaire pour permettre de lier les différents flots de données et réussir la translation de l'Épidémiologie Génétique vers l'Épidémiologie Génomique.

[6]Acide RiboNucléique

Communication scientifique

Publications associées à la thèse

Guedj, M. (2007). Interpretation Hardy-Weinberg deviations in case-controls association studies : benefits and limits *submitted in Statistics in Medecine*.

Guedj, M., Della-Chiesa, E., Picard, F., and Nuel, G. (2006a). Computing power in case-control association studies through use of quadratic approximations : application to meta-statistics. *Annals of Human Genetics*, **71**, 262–270.

Guedj, M., Robelin, D., Hoebeke, M., Lamarine, M., Wojcik, J., and Nuel, G. (2006b). Detecting local high-scoring segments : a first stage approach to genome-wide association studies. *Statistical Applications to Genetic and Molecular Biology*, **5**.

Guedj, M., Wojcik, J., Della-Chiesa, E., Nuel, G., and Forner, K. (2006c). A fast, unbiased and exact allelic test for case-control association studies. *Human Heredity*, **61**, 210–221.

Autres publications

Aschard, H., Guedj, M. and Demenais, F. (2007) A multiple-marker two-step approach for genome-wide association studies. *BMC Genetics, in press*.

Forner, K., Lamarine, M., Guedj, M., Dauvillier, J and Wojcik, J. (2007) Universal false discovery rate estimation methodology for genome-wide association studies. *Human Heredity, in press*.

Della-Chiesa, E., Guedj, M., Nuel, G. (2007) Should we combine statistics in genetic association studies. *submitted in American Journal of Human Genetics*.

Conférences internationales

European Mathematical Genetics Meeting. Talk (2007). Heidelberg, Allemagne.

International Genetic Epidemiology Society. Poster (2006). Tampa, USA.

Genetic Analysis Workshop 15. Talk (2006). Tampa, USA.

European Mathematical Genetics Meeting. Poster (2006). Cardiff, UK

International Genetic Epidemiology Society. Poster (2005). Salt Lake City, USA.

Développement logiciel

`LHiSA`	*Local High-scoring Segments for Association* `http ://stat.genopole.cnrs.fr/software/lhisa`
`fueatest`	*A Fast, Unbiased and Exact Allelic Test for association studies* `http ://stat.genopole.cnrs.fr/software/fueatest`
`kerfdr`	*A kernel-based estimation method of the Local False Discovery Rate* (en développement)

Original Paper

Human Heredity

Hum Hered 2006;61:210–221
DOI: 10.1159/000094776

Received: February 12, 2006
Accepted: June 1, 2006
Published online: July 27, 2006

A Fast, Unbiased and Exact Allelic Test for Case-Control Association Studies

M. Guedj[a, b] J. Wojcik[a] E. Della-Chiesa[a] G. Nuel[a] K. Forner[b]

[a]Statistique et Genome Laboratory, CNRS UMR 8071, Evry, France; [b]Serono Pharmaceutical Research Institute, Plan-les-Ouates, Switzerland

Key Words
Association studies · Exact test · Allelic test · Power

Abstract
Association studies are traditionally performed in the case-control framework. As a first step in the analysis process, comparing allele frequencies using the Pearson's chi-square statistic is often invoked. However such an approach assumes the independence of alleles under the hypothesis of no association, which may not always be the case. Consequently this method introduces a bias that deviates the expected type I error-rate. In this article we first propose an unbiased and exact test as an alternative to the biased allelic test. Available data require to perform thousands of such tests so we focused on its fast execution. Since the biased allelic test is still widely used in the community, we illustrate its pitfalls in the context of genome-wide association studies and particularly in the case of low-level tests. Finally, we compare the unbiased and exact test with the Cochran-Armitage test for trend and show it perfoms similarly in terms of power. The fast, unbiased and exact allelic test code is available in R, C++ and Perl at: http://stat.genopole.cnrs.fr/software/fueatest.

Introduction

Recent great progresses in the field of genotyping technologies have led to a decrease in cost and time for data production [1]. Geneticists can now consider launching large-scale genetic association studies in order to shed light on mechanisms responsible for complex human diseases and to provide opportunities for pharmaceutical companies to discover new drug targets.

Association analyzes are generally performed via a case-control design that has been preferred to family-based frameworks [2] since it is cheaper and quicker, and data are easier to collect. This approach involves unrelated individuals sampled from the same general population and clustered into two sub-populations (cases and controls) according to their disease status. When comparing the two populations, differences in terms of genotype frequencies should be due to an association between the marker and the phenotype under consideration.

Single-point strategies which treat one marker at a time are widely used as a preliminary step in the analysis. Individuals are organized into contingency tables according to their marker and disease status. Various approaches are proposed to test for association, based on either genotypes or alleles. Among them the genotypic, allelic and Hardy-Weinberg [3] chi-square tests as well as the Cochran-Armitage test for trends [4] are mainly invoked.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/06/0614–0210$23.50/0

Accessible online at:
www.karger.com/hhe

Jerome Wojcik
Serono Pharmaceutical Research Institute
14, chemin des Aulx
CH–1228 Plan-les-Ouates (Switzerland)
Tel. +41 22 706 91 61, Fax +41 22 706 99 92, E-Mail jerome.wojcik@serono.com

Table 1. Genotypic contingency table (d_i/h_i stands for the frequency of diseased/healthy individuals having i alleles A)

	aa	aA	AA	Margins
Case (diseased)	d_0	d_1	d_2	N_d
Control (healthy)	h_0	h_1	h_2	N_h
Margins	N_0	N_1	N_2	N

Table 2. Corresponding allelic contingency table

	a	A	Margins
Case	$2d_0 + d_1$	$2d_2 + d_1$	$2N_d$
Control	$2h_0 + h_1$	$2h_2 + h_1$	$2N_h$
Margins	$2N_0 + N_1$	$2N_2 + N_1$	$2N$

In this article, we focus on the approach based on allele frequencies to test for association. It has been shown to be unadapted when alleles are not paired independently [5] but it remains widely used. As an alternative we propose an unbiased test based on the exact computation of the p value. Since highthroughput data require to perform a lot of tests, we have paid attention to the speed of execution by optimizing the implementation to make each exact test the less time consuming possible. Then an application on experimental data gives an idea about the real impact of using the biased test in the context of genome-wide association studies. The Cochran-Armitage test for trends has also been suggested as an alternative to the biased alellic test. To guide the choice between the two alternatives, we propose to study their respective powers focusing on situations where genotype frequencies do not comply with Hardy-Weinberg equilibrium (HWE).

Methods

Testing for Association

Let us denote X, a case-control sample. Testing for association requires first to consider a null hypothesis (H_0) used to test a particular distribution of the observations. To do so, we consider a statistic defined as a function of the observations: $S = f(X)$, and carefully chosen such that S grows when H_0 is less likely. Using the distribution of S under H_0, we can find a threshold (t_α) such as $\alpha = \mathbb{P}_{H_0}(S \geq t_\alpha)$, where level α can be set to 5% for instance. The significance of an observed value (S^{obs}) is termed p value and corresponds to the probability to observe data as or more extreme than the observation if H_0 were true: p value $= \mathbb{P}_{H_0}(S \geq s^{\text{obs}})$.

The Biased Allelic Test

Two strategies exist to tabulate data from genetic case-control studies: the first classifies cases and controls according to their genotypes (table 1) and the second according to their alleles (table 2).

Allelic association tests aim at detecting significant differences in allelic frequencies between cases and controls testing the null hypothesis of no association with the disease. Alleles are supposed to be sampled from the case and control populations with probabilities p_d and p_c corresponding to the proportions of the susceptibility allele in cases and controls respectively. Under H_0, alleles are sampled from the same general population and p_d and p_c are equal to p. To test this hypothesis we classically consider the statistic:

$$Z_A = \frac{\hat{p}_d - \hat{p}_c}{\sqrt{N\hat{p}\left(1-\hat{p}\right)}},$$

with

$$\hat{p}_d = \frac{2d_2 + d_1}{2N_d}, \hat{p}_c = \frac{2c_2 + c_1}{2N_c} \text{ and } \hat{p} = \frac{2N_2 + N_1}{2N},$$

or the equivalent Pearson statistic applied to the allelic table:

$$S_A = \left(Z_A\right)^2 = \frac{2N \cdot \left(\left(2d_2 + d_1\right)\left(2c_0 + c_1\right) - \left(2d_0 + d_1\right)\left(2c_2 + c_1\right)\right)^2}{2N_d 2N_c \cdot \left(2N_2 + N_1\right)\left(2N_0 + N_1\right)}.$$

Focusing on S_A, the p value can be calculated using two testing procedures: (i) Under H_0 and when data are in accordance with Cochran's conditions (each expected cell count greater than 5) [6], the distribution of S_A can be asymptotically well-approximated by a one degree-of-freedom (df) chi-square distribution. This comes down to the well-known chi-square test on contingency tables applied, in our case, to the 2×2 allelic table. (ii) An exact procedure also exists: the probability to observe a table conditionally to the margins is given by the hypergeometric distribution. Hence enumerating the whole set of 2×2 tables in accordance with the observed margins of the allelic table, the exact p value comes down to the sum of the probabilities to observe 2×2 tables for which S_A is at least as great as s_A^{obs}.

Strategies based on allelic contingency tables may appear appealing since the sample size is doubled in comparison with the genotypic approach. However the validity of this test has been recently discussed and some authors recommend against its use [5, 7]. As described below, the null hypothesis makes two assumptions:

$$H_0^{(\text{biased})}: \left\{ \begin{array}{ll} 1. & p_d = p_c \\ 2. & \text{Alleles are sampled independently} \end{array} \right\}.$$

Although genotypes are sampled independently from cases and controls, it is however clear that alleles are not: in the allelic contingency table, each individual contributes to two observations so alleles are actually sampled by two at a time. As a result, the way alleles are paired (corresponding to a departure from HWE) deviates S_A from its wrongly supposed distribution under H_0 and hence dramatically biases the p value.

The Cochran-Armitage Test for Trends

Instead, Sasieni recommended to use the Cochran-Armitage test for trends [5]. It has been shown to be asymptotically equivalent to the allelic test when the combined population is in HWE (cases and controls) but makes no assumption about allele matching. It is based on the genotypic contingency table. Genotypes are sampled from the case and control populations with probabilities $(p_{d_0}, p_{d_1}, p_{d_2})$ for cases and $(p_{c_0}, p_{c_1}, p_{c_2})$ for controls, which is more consistent with reality. The null hypothesis for the Cochran-Armitage test is no trend, which means that the proportions p_{d_i} are the same for all genotypes:

$$H_0^{(\text{trend})} : \{p_{d_0} = p_{d_1} = p_{d_2}\}.$$

In practice, the trend statistic S_T measures a linear trend in proportions weighted by a dose effect score x_i associated to each column of the contingency table. When x_i quantifies the number of high-risk allele, this test is equivalent to the score test in the logistic regression model where each SNP is coded according to its count of high-risk allele. In our case:

$$S_T = \frac{N \cdot \left[N \cdot (d_1 + 2d_2) - N_d \cdot (N_1 + 2N_2)\right]^2}{N_d N_h \cdot \left[N \cdot (N_1 + 4N_2) - (N_1 + 2N_2)^2\right]} \underset{H_0}{\sim} \chi^2$$

An Unbiased and Exact Allelic Test

As a more natural alternative to the biased allelic test, we propose to perform an unbiased test still based on the allelic Pearson statistic (S_A) and on the sampling of genotypes instead of alleles taken independently.

$$H_0^{(\text{unbiased})} : \{p_d = p_c\}.$$

Following this direction, Schaid and Jacobsen [8] found out that using the classical but biased allelic test alters the type I error-rate in a predictable manner and proposed a correction on the chi-square approximation. Our alternative relies on an exact computation of p values and can be thus termed as an unbiased and exact test. The principle is quite similar to the biased and exact procedure described below with the difference that instead of enumerating the 2 × 2 tables, we enumerate all the 2 × 3 tables in accordance with the observed margins of the genotypic table. In such a way, alleles are not considered individually, H_0 takes the allele matching into account and makes only the assumption of no association. Probabilities are computed using a multiple hypergeometric distribution and as previously, the exact p value comes down to the sum of the probabilities to observe 2 × 3 tables for which S_A is at least as great as s_A^{obs}.

Implementation

Here we propose an accurate description of the implementation for this test. Divided in three main steps, it intends to avoid most of errors due to computation.

Enumerating the Genotypic Tables

Conditionally to the margins, 2 × 3 contingency tables have two degrees of freedom. Consequently a table is completely determined by setting two of its elements, for example d_0 and d_1 that we refer now to as i and j to stress out that they are integer variables (table 3). These variables can take a range of values which depends on the margins:

Table 3. $T_{i,j}$ the 2 × 3 genotypic table with respect to the values of i, j and the margins

	aa	*aA*	*AA*	Margins
Case	i	j	$N_d - i - j$	N_d
Control	$N_0 - i$	$N_1 - j$	$N_h - N_0 - N_1 + i + j$	N_h
Margins	N_0	N_1	N_2	N

- i takes all integer values in:

$$[i_{\min}, \ldots, i_{\max}] \overset{def.}{=} \left[\max(0, N_1 - N_h), \ldots, \min(N_1, N_d)\right],$$

- for a given value of i, j can take all integer values in:

$$[j_{\min}(i), \ldots, j_{\max}(i)] \overset{def.}{=} \left[\max(0, N_d - N_3 - i), \ldots, \min(N_2, N_d - i)\right].$$

Determining these intervals of variation allows us to enumerate the whole set of 2 × 3 tables in accordance with the margins of the observed genotypic table.

Comparing Pearson Statistics

This issue can appear trivial but can generate errors and lead to inaccurate results: on a computer, real numbers are stored in floating-point format for which the arithmetic is not exact [9]. As a result, two scores (s_1, s_2) that are supposed to be mathematically equal can show different computational representations so that the result of the logical expression $s_1 \geq s_2$ is unpredictable. In our case, it is possible to circumvent this problem by coming back to integer arithmetic that is mercifully exact on computers.

To do so we first outline a simplified expression for the Pearson statistic on 2 × 2 tables. By definition, for a given allelic table

a b
c d

this statistic is usually expressed as:

$$S_A = \frac{(a - \bar{a})^2}{\bar{a}} + \frac{(b - \bar{b})^2}{\bar{b}} + \frac{(c - \bar{c})^2}{\bar{c}} + \frac{(d - \bar{d})^2}{\bar{d}}$$

where $\bar{a}, \bar{b}, \bar{c}, \bar{d}$ are the expected values of a, b, c and d under H_0, i.e. the products of the margins divided by the total number $n = a + b + c + d$. For instance

$$\bar{a} = \frac{(a+b)(a+c)}{n}.$$

An interesting property is that all the numerators are equal. For example

$$\left(b-\bar{b}\right)^2 = \left(b - \frac{(b+a)(b+d)}{n}\right)^2 = \left(\frac{nb-(a+b)(n-(a+c))}{n}\right)^2 = (a-\bar{a})^2.$$

The expression of s_A comes down to:

$$s_A = \left(\frac{1}{\bar{a}} + \frac{1}{\bar{b}} + \frac{1}{\bar{c}} + \frac{1}{\bar{d}}\right)(a-\bar{a})^2 = \lambda(a-\bar{a})^2 \quad (1)$$

where $\bar{a}$ and λ only depend on the margins and hence are constant. We take advantage of this property to compare statistics by the way of integers: for two tables $T_1 = (a_1, b_1, c_1, d_1)$ and $T_2 = (a_2, b_2, c_2, d_2)$ with respective statistics s_1 and s_2, it follows that:

$$s_1 \geq s_2 \Leftrightarrow |na_1 - n\bar{a}| \geq |na_2 - n\bar{a}|,$$

which is computationally interesting since the statement $|na_1 - n\bar{a}| \geq |na_2 - n\bar{a}|$ only involves integers.

Computing the Probability of a 2 × 3 Table

The probability of a 2 × 3 table T under H_0 is given by the multiple hypergeometric distribution:

$$p(T) = \frac{N_d!N_h!N_1!N_2!N_3!}{N!d_0!d_1!d_2!h_0!h_1!h_2!}$$

Numbers under consideration are so huge that the straightforward approach is not computationally feasible. Instead we use the log-factorial [9] and exponential functions to compute approximates.

Let us note $\mathrm{LF}(k) = \log(k!) = \sum_{i=1}^{k} \log(i)$, so that $\exp(\mathrm{LF}(k)) \sim k!$. Thus:

$$p(T) \simeq \exp(\Sigma\ \mathrm{LF}(k) - \Sigma\ \mathrm{LF}(k')) \quad (2)$$

for $k \in \{N_d, N_h, N_1, N_2, N_3\}$ and $k' \in \{N, d_0, d_1, d_2, h_0, h_1, h_2\}$. The accuracy of these approximations is greater than 1.10^{-10}.

Optimizations

In terms of implementation, the main difference between the exact biased and unbiased tests relies on the number of tables to enumerate. In most cases, considering 2 × 3 contingency tables in accordance with the data is likely to increase the execution time. The dimension of available data requires to compute thousands of tests. In order to speed up analyzes, we optimized our implementation, working on (i) a sensitive reduction of the number of tables to enumerate and (ii) a recursive relation between table probabilities $p(T)$. Interested readers can refer to the Appendixes 1 and 2 for more details.

Algorithm 1. Pseudo-code of the basic implementation

```
compute all margins N_d, N_h, N_1, N_2, N_3, N and n = 2N
compute the observed allelic count a = 2d_0 + d_1 and nā
compute the observed threshold t = |na − nā|
let p value = 0
for i = max(0, N_1 − N_h) to min(N_1, N_d)
for j = max(0, N_d − N_3 − i) to min(N_2, N_d − i)
if |n(2i + j) − na| ≥ t then
compute p(T_{i,j})
p value = p value + p(T_{i,j})
output the p value
```

Application

We use an application on real data to assess the consequences of using the biased test, first in terms of p value deviation from the unbiased p value (qualitative results), but also in terms of effective impact on predictions (quantitative results).

Data

We applied the biased and unbiased exact tests on genome-wide data concerning the multiple sclerosis. The data set consists in 66,990 SNPs. DNA from 279 patients and 301 controls were genotyped with this marker set using the 100K Affymetrix chip. The algorithm used for making genotype calls has been previously described by Affymetrix [10].

Notation

In what follows, p_u and p_b refer to the unbiased and biased exact allelic p values respectively. Let p be the frequency of the allele of susceptibility, p_{HW} the p value for the HW test on combined population and f the coefficient of consanguinity quantifying a deficit ($f > 0$) or excess ($f < 0$) of heterozygotes.

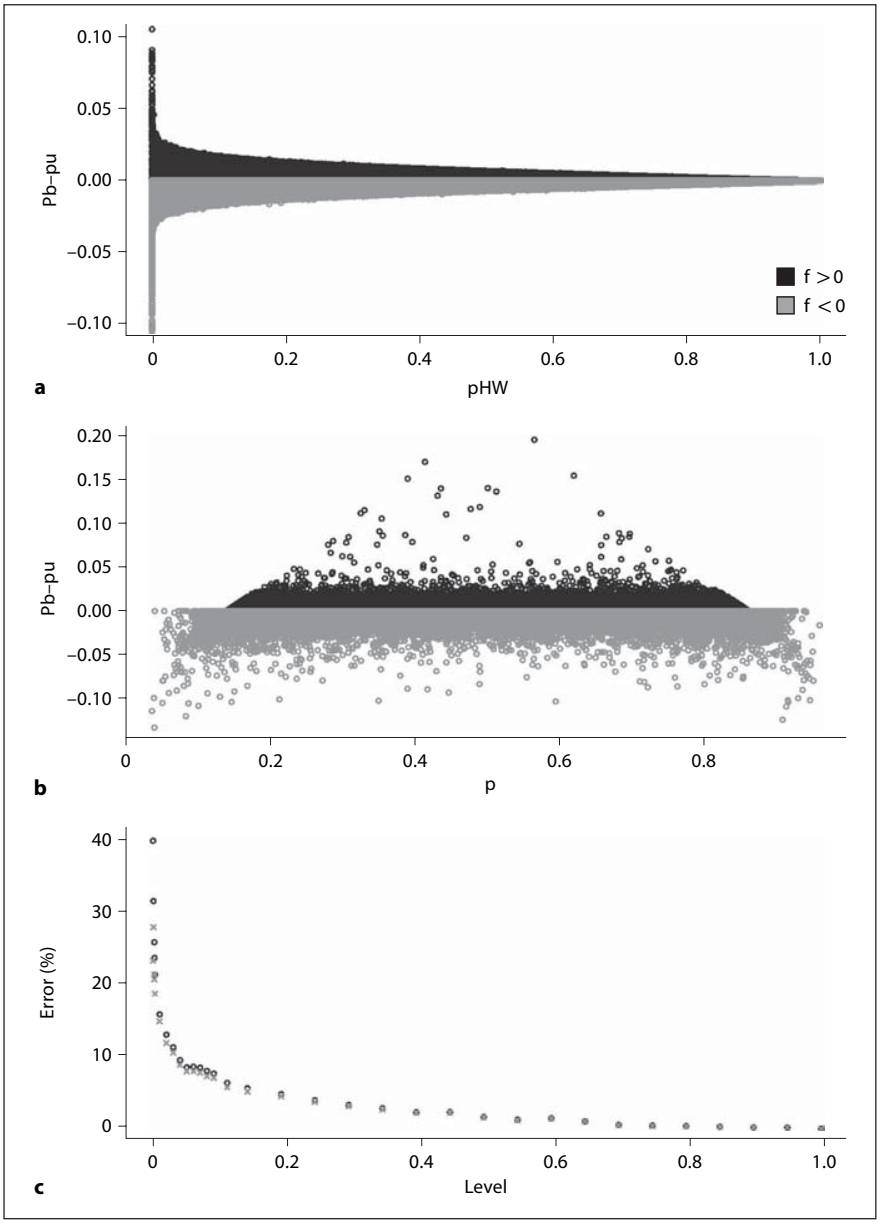


Fig. 1. Deviation of p_b from p_u: **a** This graphic represents the absolute error made using p_b to approximate p with respect to the strength of departure from HWE. **b** It outlines the dependence between the absolute error and the frequency of the allele of susceptibility (p). **c** The percentage of errors is estimated by the ratio of total false positives and negatives made by the use of the biased test on the count of positives predicted by the unbiased exact test. It is represented with respect to the level of the test for the whole set of SNPs (○) and this same set excluding SNPs that fail the HWE test (×).

Qualitative Results

First, figure 1 displays the absolute error made using p_b to approximate p_u with respect to the strength of deviation to HWE and the allele frequency. As expected, the amount of deviation is directly proportional to the departure from HWE (fig. 1a). The way alleles are paired deviates p_b in a predictive direction: a deficit of heterozygotes increases p_b and at the same time the rate of false positives. By opposition an excess of heterozygotes decreases p_b and so increases the conservatism of the test. These observations are also illustrated in the examples in table 4. It is clear that p_u and p_b are strictly identical when the combined population is in HWE and that the absolute error increases with the amount of departure from the equilibrium.

Figure 1b represents the link that exists between allele frequencies (p) and the deviation of p_b. In the case of a deficit of heterozygotes, there is absolutely no visible de-

Table 4. Biased, unbiased and Monte-Carlo p values for five representative examples: each example varies in the strength of departure from the Hardy-Weinberg equilibrium and in the direction of departure

p_{HW}	f	Genotypic table			p_u	p_b	p_{MC}
1	= 0	29	117	124	0.571	0.571	0.571
		34	135	128			
0.5	> 0	26	109	113	0.326	0.318	0.326
		28	111	144			
0.5	< 0	97	113	26	0.347	0.353	0.347
		128	123	28			
0.05	> 0	10	77	140	0.310	0.288	0.310
		24	82	161			
0.05	< 0	155	94	9	0.228	0.249	0.229
		171	90	5			

p_{MC} was computed based on 10^6 simulations.

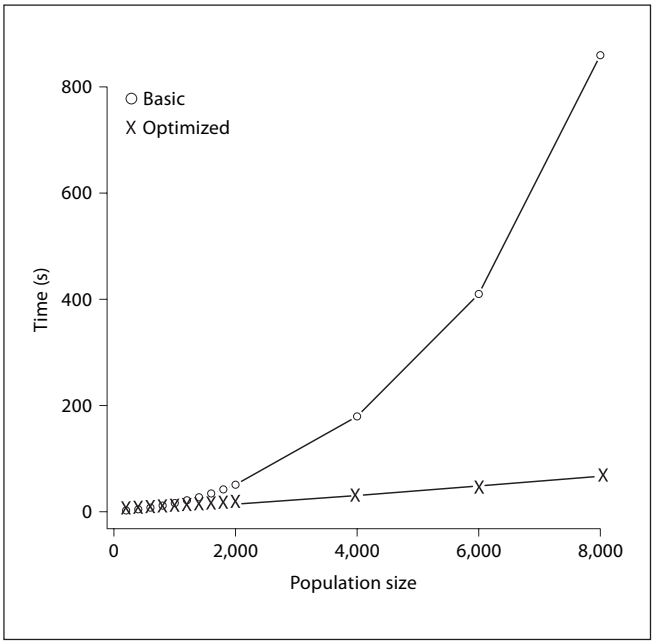


Fig. 2. Computational time: the time of execution is assessed for the basic and the optimized algorithms on 10,000 simulated SNPs and increasing sizes of population.

pendence but in the case of an excess, the maximum amount of possible deviation is proportional to the quantity $p(1 - p)$ and hence maximal for $p = 0.5$ and minimal at the bounds $p = 0$ and $p = 1$.

Quantitative Results

Even more interesting is the percentage of erroneous predictions made with the biased allelic test (fig. 1c). It is quite substantial and increases as the level of the test decreases up to 9, 16, 30 and 40% for respectively a 5, 1, 0.1 and 0.01% level respectively. Since most of researchers would exclude SNPs that clearly fail HWE in controls [11], we also computed these erroneous prediction rates excluding SNPs for which the HWE test p values for controls does not exceed the 5% significance level (when taking into account multipletesting using the Benjamini-Hochberg procedure [12]). In this case, these rates are less important but remain substantial, up to 28% for a test level of 0.01.

Validation

Our implementation intends to avoid errors of computation. To be convinced of the exactness of our procedure, we compared our p values with p values calculated with an infinite precision, which requires an important time of execution. Both appear to be identical (data not shown). Moreover, we compared our p values with p values calculated with Monte-Carlo simulations (table 4). The distribution of a statistic under H_0 can be empirically deduced via a set of simulations, permuting at each iteration the case and control labels. Considering the precision due to the number of simulations performed, the situations tested show that exact and empirical p values are not significantly different.

Computational Time

We ran some benchmarks on simulated data to evaluate the gain made by the proposed optimizations. The machine used is an Itanium 2 (64 bits, 1.6 GHz) running GNU/Linux. We performed the unbiased exact test with the basic and optimized implementations on 10,000 markers for different sizes of population (fig. 2). The time of execution is strongly linked to the power of the computer but we can notice the great advantage of the optimized implementation. For large populations, the gain grows with the population size up to 6 and 12 times faster for data implicating 4,000 and 8,000 individuals respectively.

Unbiased and Trend Tests: Comparative Assessment

Two alternatives are proposed to the biased allelic test based on two different statistics (s_A and s_T) and testing for two slightly different null hypotheses, as described previ-

ously. These two approaches have been compared in a power study based on Monte-Carlo simulations: as long as it is possible to generate a case-control sample X under an alternative hypothesis (H_1), it is very easy to get an estimation of the power $\pi(\alpha) = \mathbb{P}_{H_1}(s \geq t\alpha)$. We first draw B samples denoting $x^{(i)}$ the i-th sample. From this sample we get B statistics $s^{(1)}$, ..., $s^{(B)}$ from which we compute the estimate:

$$\hat{\pi}(\alpha) = \frac{\#\left\{s^{(i)} \geq t_a\right\}}{B}.$$

This method is widely used in the field of statistical genetics when alternative distributions are hard to derive analytically [13]. But computing $\hat{\pi}(\alpha)$ requires first to define a genetic model. Let a bi-allelic disease susceptibility locus (DSL) with its two possible alleles (the allele of susceptibility A and a), p be the frequency of the allele A and r_0, r_1 and r_2 the frequencies of the genotypes aa, aA and AA respectively. According to Wright's model [14], we have: $r_2 = p^2 + fp(1 - p)$, $r_1 = 2p(1 - p) - 2fp(1 - p)$ and $r_0 = (1 - p)^2 + fp(1 - p)$ with f the coefficient of consanguinity defined previously. Now we introduce the prevalence of the disease (K_p), and the penetrances (f_i) associated to each genotypes (i). Considering the relative risks (RR_i) such as $RR_i = f_i/f_0$ for $i = 1$ or 2, we define the four main modes of inheritance (MOI) underlying the mode of action of the DSL on the disease: recessive ($RR_1 = 1$), multiplicative ($RR_1 = \sqrt{RR_2}$), additive ($RR_1 = (RR_2 + 1)/2$) and dominant ($RR_1 = RR_2$). From these parameters, we can easily derive $f_0 = K_p/(r_0 + RR_1 \cdot r_1 + RR_2 \cdot r_2)$, $f_1 = RR_1 \cdot f_0$ and $f_1 = RR_1 \cdot f_0$ and finally genotype frequencies in case and control populations:

$$\left(p_{d_0}, p_{d_1}, p_{d_2}\right) = \left(\frac{f_0 \cdot r_0}{K_p}, \frac{f_1 \cdot r_1}{K_p}, \frac{f_2 \cdot r_2}{K_p}\right)$$

$$\left(p_{c_0}, p_{c_1}, p_{c_2}\right) = \left(\frac{(1-f_0) \cdot r_0}{1-K_p}, \frac{(1-f_1) \cdot r_1}{1-K_p}, \frac{(1-f_2) \cdot r_2}{1-K_p}\right)$$

In such a context, H_0: $\{RR_2 = 1\}$ and H_1: $\{RR_2 > 1\}$.

Simulations

Simulations are performed considering three distinct sets. The power is estimated first with respect to the susceptibility allele frequency (p), then according to the coefficient of consanguinity (f) introduced in the general population and finally we investigate the effect due to the variation of subpopulation sizes (N_d and N_h). All situations are considered for a prevalence $K_p = 0.05$, the four MOI and two strengths of disease implication ($RR_2 = 1.5$ and $RR_2 = 2$). Each estimate is done on the basis of 10,000 simulations.

Results

Main results are summarized in figures 3 and 4. Full results are compiled in supplementary data available upon request. With the relative risks considered, the additive and multiplicative models present actually close values of RR_1 and hence very similar results. In the case where p varies (fig. 3), the power of the trend and the unbiased allelic tests are not significantly different. They are also similar to the biased allelic test which is consistent with the fact the three tests are strictly identical when HWE holds in the combined population. However differences can still be observed since a departure from the equilibrium in cases can be generated by the disease.

When the equilibrium does not hold in the general population ($f \neq 0$, fig. 3), both trend and unbiased allelic tests show comparable power with each MOI, increasing along with f. We also represented the power of the biased allelic test (dotted) although there is no sense to take it in consideration. Previous conclusions about the deviation of the type-I error rate with respect to f corresponds here to an artificial decrease ($f < 0$) and increase ($f > 0$) of power. Despite their comparative efficiency in all situations studied, the trend and unbiased allelic tests however show small but significant differences. Under a dominant model, the trend test tends to be more powerful than the unbiased allelic test which, by opposition, appears more powerful under a recessive model. These differences are clearer when the population contains a high proportion of heterozygotes ($f < 0$) and increase with the ratio N_d/N_h (fig. 4). In addition, investigating the potential influence of the sample size, results outline that an augmentation of N clearly accentuates the bias introduced by the biased allelic test.

On our real dataset, the biased p value distribution appears closer to the unbiased one than to the trend one (fig. 5a–c). This is logical since the biased and unbiased tests are based on the same statistic (s_A) and almost the same H_0 hypothesis. In practice, differences in term of findings are not important between the unbiased and the trend test, up to 8%, with a mean of 4% along with the level of the test (fig. 5d).

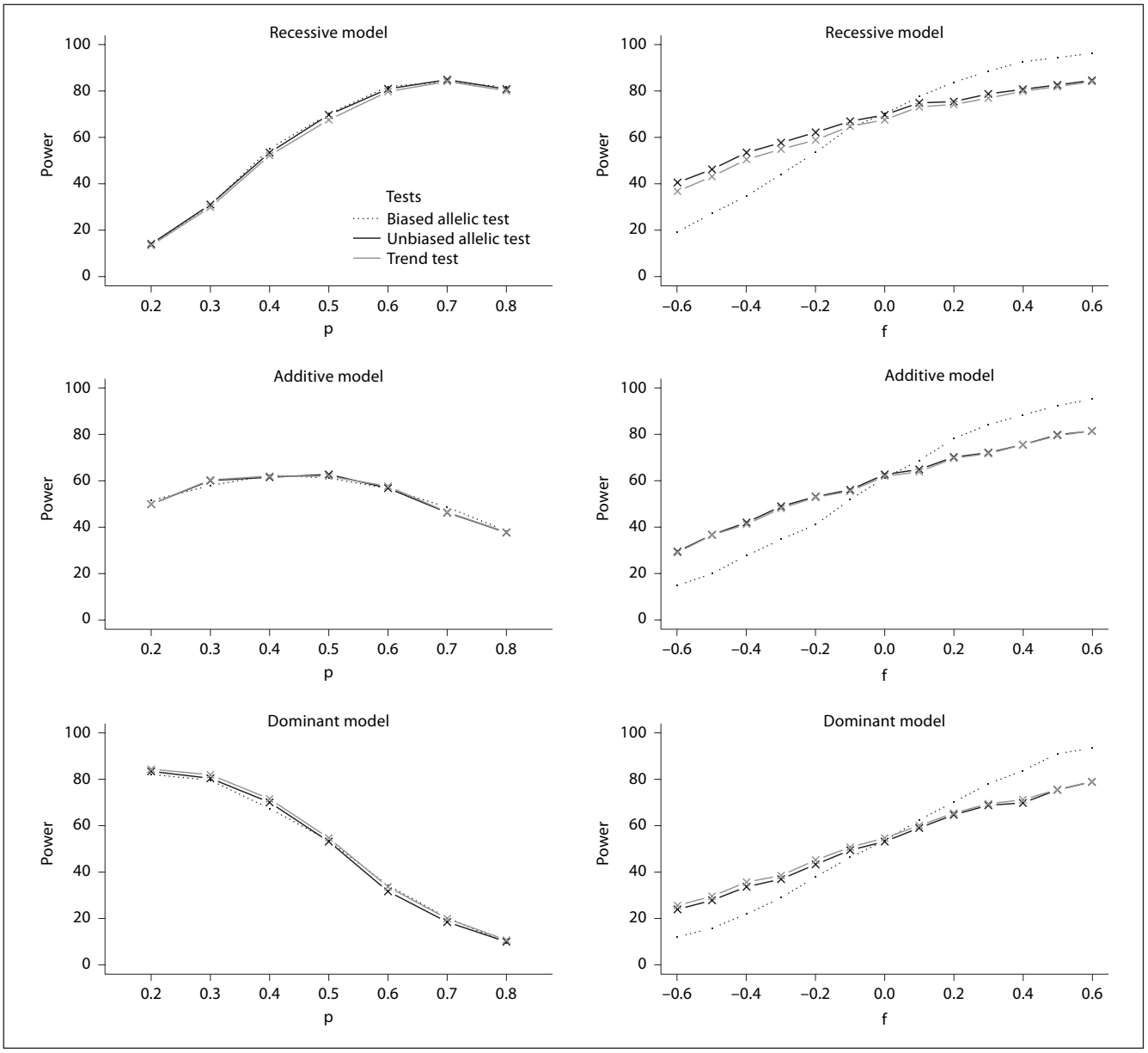


Fig. 3. Power estimation with respect to p and f ($K_p = 0.05$, $RR_2 = 1.5$ and $n_D = n_C = 500$).

Discussion

With the work of Sasieni, it is today well-known that the classical allelic test which assumes an independence of alleles is biased by the allele matching [5]. The solution we developed here constitutes an unbiased and exact test to compare allele frequencies. It has the advantage of making no assumption about the way alleles are paired and unlike asymptotic tests, its validity is not constraint to the filling of the contingency tables. Based on real genomewide data, our results clearly outline that the strength of deviation directly depends on the importance of the departure from HWE. Moreover, the direction in which the p value deviates is linked to the way the equilibrium is not respected. The sample size also appears to have an effect on the width of the bias as well as the allele

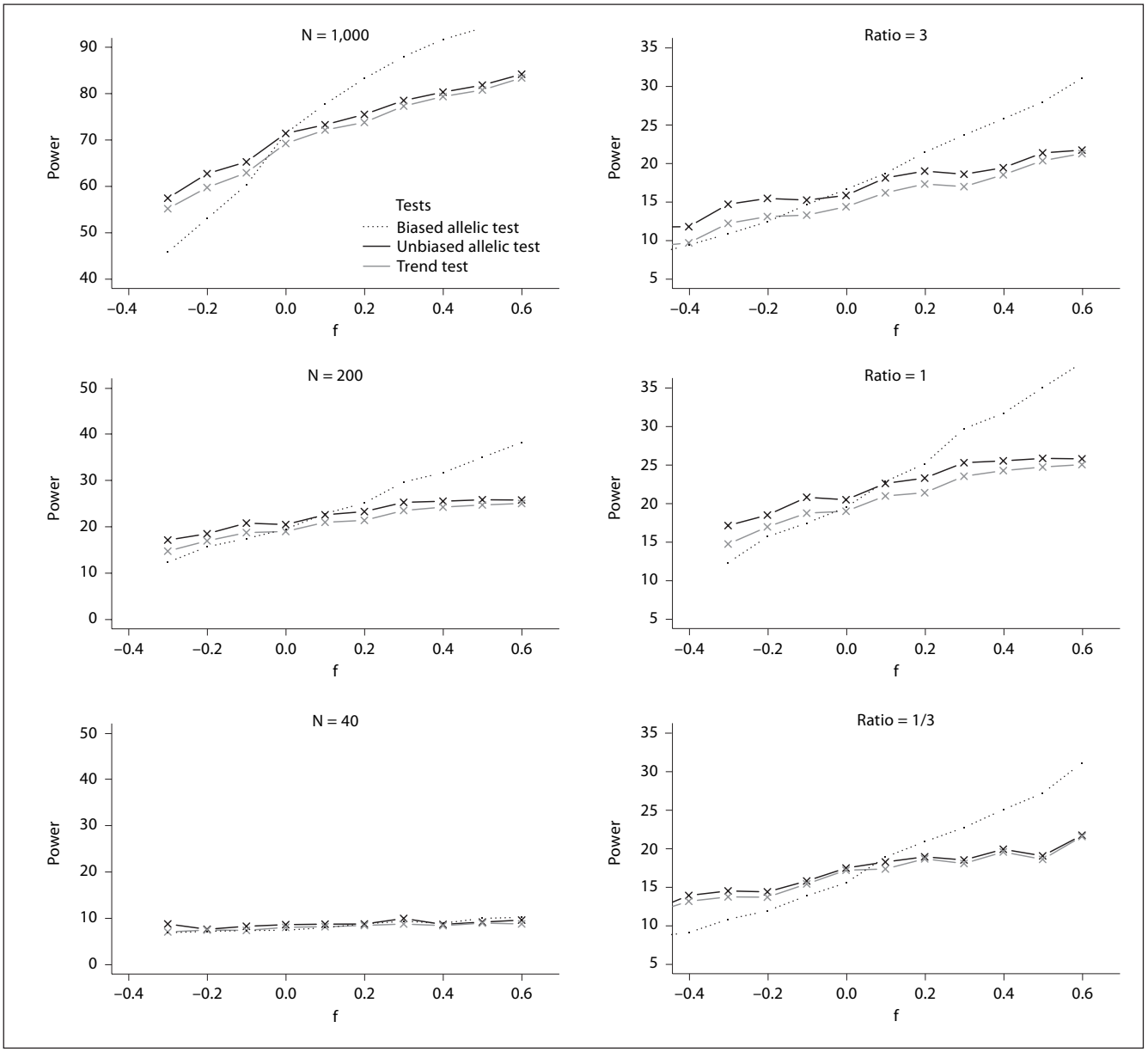


Fig. 4. Power estimation with respect to *f*, *N* and the ratio N_d/N_c ($K_p = 0.05$, $p = 0.5$, $RR_2 = 1.5$, $RR_1 = 1$.)

frequency in the case of an excess of heterozygotes. Some of these observations actually confirm and illustrate on genome-wide real data results analytically deduced by Schaid and Jacobsen [8]. Finally, we show that this bias can have a substantial impact on predictions that increases as the level of the tests decreases.

In practice, association studies are not working in a favorable situation with regard to our conclusions: (i) the threshold chosen to decide for the acceptance or the rejection of the H_0 hypothesis is generally low, commonly set to 5 or 1% and even lower when one takes the multiple-testing into account; (ii) recent studies aim to detect common susceptible alleles responsible for common diseases, and (iii) involve hundreds of subjects. As a result, if the impact on true findings is not easy to assess in real data, it remains potential and likely to be substantial. One usu-

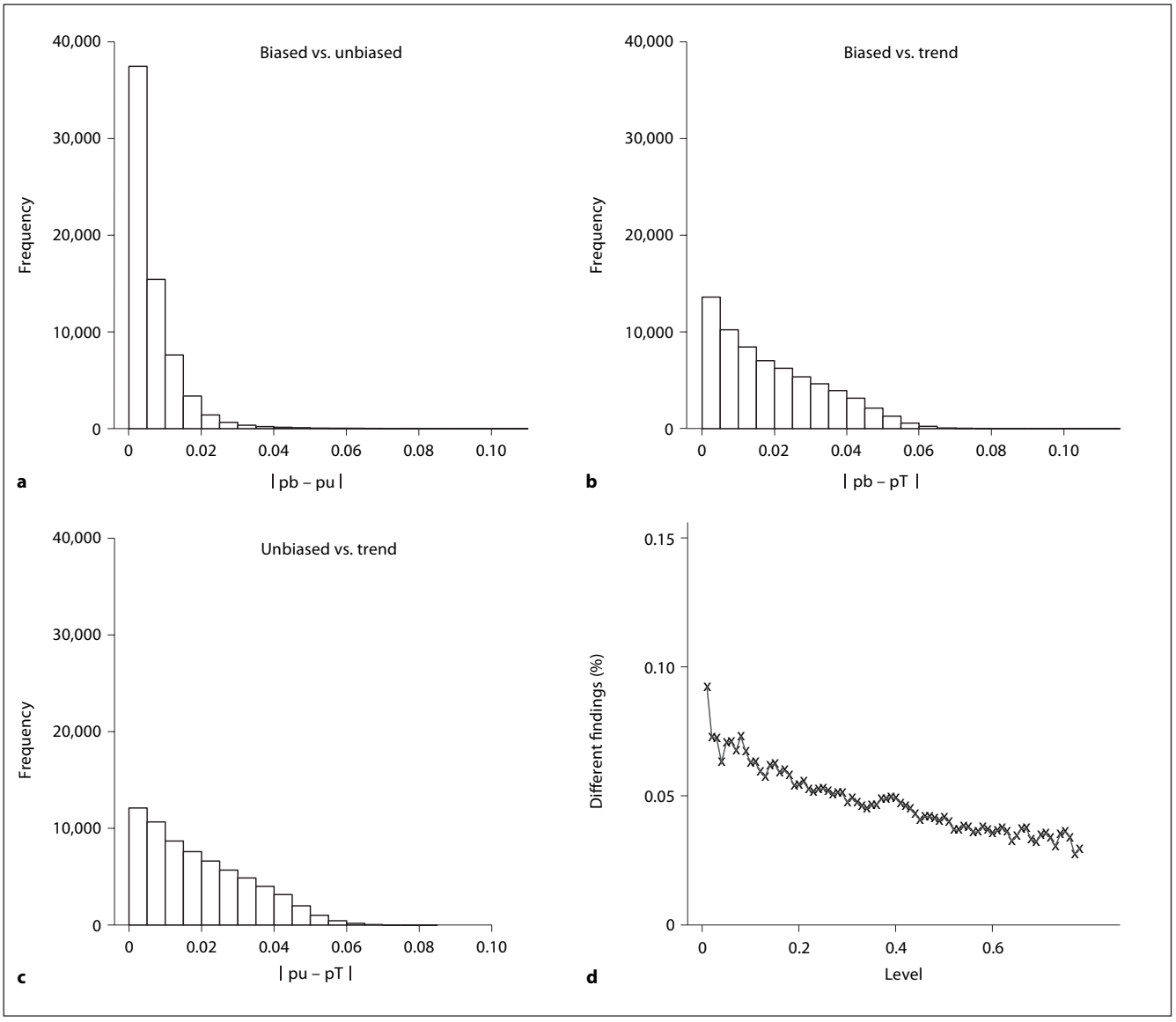


Fig. 5. a–c Histograms of absolute differences between the p-values of the biased, unbiased and trend tests. **d** It represents the percentage of different findings between the biased and trend test with respect to the level of the test.

al way to deal with HWE violations is to discard markers that fail HWE in controls with the argument that disequilibrium may be indicative of genotyping errors [11]. If such a strategy manages to reduce the percentage of erroneous predictions, it still represents an important amount of errors. In addition, HWE may be violated because of genotyping error but not only: chance, nonrandom mating, inbreeding, genetic drifting, selection, population stratification or combination of these reasons are mainly evoked and also susceptible to introduce a statistical bias on allele-based contrast [15]. Because the origin of HW disequilibrium can not be ensured, unbiased tests should reasonably be applied on all markers, with additional attention on interpretations when HWE does not hold in controls. And besides the impact on predictions, the exactness of *p* values can be really important when one applies methods based on *p* value distributions such as FDR-based approaches [16].

Compared to the trend test, the unbiased allelic test presents comparable power under various alternatives

and results on real data are quite similar. Consequently the choice between the trend or the unbiased allelic tests should rather be based on practical reasons. As underlined by Knapp, the allelic statistic (s_A) appears more intuitive than the trend statistic (s_T) to the genetics community [17]. Moreover, we see the unbiased test as a more natural alternative to the biased test and we have developped the unbiased test in various languages to support its diffusion. With the recent accumulation of data, genetic studies require the simultaneous treatment of a large amount of markers (up to 500,000). Besides the statistical problems that the analysis raises (multiple-testing, curse of dimensionality), the time of execution is also an important issue. Exact testing procedures are time consuming and so we found relevant to improve the basic implementation in order to accelerate the algorithm execution. Using the optimized implementation is clearly advantageous for the treatment of large populations. Such computational considerations, that can be easily extended to other contexts, along with the current power of computers make the use of exact tests reasonable.

A major issue in genetic studies is to observe replications in independent populations to ensure the reliability of the results but they are generally difficult to obtain [18, 19]: with the complexity of the diseases under investigation, lack of power, multiple-testing [12], genotyping errors [20] and population stratification [21, 22] are generally invoked to explain the difficulty in replicating findings. In such a context it is important to avoid errors due to the choice of the method of analysis. The procedure we propose meets this goal. Unbiased, it avoids false predictions produced by the use of the biased allelic test. Exact, it avoids additional errors due to asymptotic approximations and computations. Fast and available under different versions, its execution is easy and adapted to very large data sets: http://stat.genopole.cnrs.fr/software/fueatest.

Acknowledgements

We would like to thank Serono for generating data and more particularly Hiroaki Tanaka as well as Bernard Prum and Franck Picard for fruitful discussions.

Appendix 1: Optimized Implementation

Reducing the Number of Tables to Enumerate

To compute p values, we can restrain the enumeration to tables that present a better/equal or worse statistic than the observed one, termed as 'better' and 'worse' tables respectively:

$$p\text{ value} = \sum_{\text{better tables}} p(T) = 1 - \sum_{\text{worse tables}} p(T)$$

Considering the smallest set of tables (instead of the whole one) decreases the time of execution but also the accumulation of imprecision. Since in genome-wide data most of the variables are likely to be under H_0, it may generally exists less worse tables than better tables. Therefore we will now focus on obtaining this subset of worse tables. The analog reasoning can be proposed for better tables.

For a given value of i, let $J_w(i)$ be the set of j values corresponding to worse tables. Defining floor(x) as the highest integer $\leq x$ and ceil(x) as the lowest integer $\geq x$, it can be shown that:

$$J_w(i) = \left[\text{floor}\left(\frac{(n\bar{a}-t)}{n}\right) - 2i + 1,\ \text{ceil}\left(\frac{(n\bar{a}+t)}{n}\right) - 2i - 1\right] \cap \left[j_{\min}(i),\ j_{\max}(i)\right]$$

As a result, a given value of i is obviously useless if $J_w(i)$ is empty for this value. So we can also describe the useful set I_u of values for i for which $J_w(i) \neq \emptyset$. See Appendix 2 for more details.

Recursive Relation between Table Probabilities under H_0

In principle, the computation of the probability $p(T_{i,j})$ requires a call to the exponential function (see Relation 2). Noticing the recursive relation that exists between $p(T_{i,j})$ and $p(T_{i,j+1})$, it is possible to use this function only once per i value:

$$p\left(T_{i,j+1}\right) = \frac{(N_d - i - j)(N_1 - j)}{(j+1)(N_h - N_0 - N_1 + i + j + 1)} \times p\left(T_{i,j}\right)$$

To save some more computational time, we can pre-calculate LF(t) for $t \in [0, ..., N]$ as well as $K = \Sigma\text{LF}(k) - \text{LF}(N)$ for $k \in \{N_d, N_h, N_1, N_2, N_3\}$ that is constant through the algorithm execution.

Appendix 2: Determining the Set I_u

From appendix 1 we have $J_w(i) = [\alpha - 2i, \beta - 2i] \cap [j_{min}(i), j_{max}(i)]$ with

$$\alpha = \text{floor}\left(\frac{n\bar{a}-t}{n}\right) - 1$$

and

$$\beta = \text{ceil}\left(\frac{n\bar{a}-t}{n}\right) + 1.$$

In the very particular case where $\alpha > \beta$, we can see that $J_w(i) = \emptyset$ $\forall i$, meaning that there is no 'worse' table and that the p value is exactly 1. So in the following we will suppose that $\alpha \leq \beta$:

$$J_w(i) \neq \emptyset \Leftrightarrow (\alpha - 2i \leq j_{max}(i)) \text{ AND } (j_{\min}(i) \leq \beta - 2i),$$

but

$$\alpha - 2i \le j_{\max}(i) \Leftrightarrow \alpha - 2i \le \min\left(N_1, N_d - i\right),$$
$$\Leftrightarrow \left\{ \begin{array}{l} \dfrac{\alpha - N_1}{2} \le i \le N_d - N_1 \\ i > N_d - N_1 \quad \text{AND} \quad i \ge \alpha - N_d \end{array} \right\},$$
$$\Leftrightarrow \left\{ \begin{array}{l} \text{ceil}\left(\dfrac{\alpha - N_1}{2}\right) \le i \le N_d - N_1 \\ \max\left(\alpha - N_d, N_d - N_1 + 1\right) \le i \end{array} \right\},$$

and

$$j_{\min}(i) \le \beta - 2i \Leftrightarrow \max\left(0, N_d - N_2 - i\right) \le \beta - 2i,$$
$$\Leftrightarrow \left\{ \begin{array}{l} i \le N_d - N_2 \quad \text{AND} \quad i \le \beta - N_d - N_2 \\ i > N_d - N_2 \quad \text{AND} \quad i \le \text{floor}\left(\dfrac{\beta}{2}\right) \end{array} \right\},$$
$$\Leftrightarrow \left\{ \begin{array}{l} i \le \min\left(N_d - N_2, \beta - N_d + N_2\right) \\ N_d - N_2 + 1 \le i \le \text{floor}\left(\dfrac{\beta}{2}\right) \end{array} \right\},$$

hence

$$i \in I_u \Leftrightarrow \left\{ \begin{array}{c} \text{ceil}\left(\dfrac{\alpha - N_1}{2}\right) \le i \le \min\left(N_d - N_2, \beta - N_d + N_2, N_d - N_1\right) \\ \max\left(\text{ceil}\left(\dfrac{\alpha - N_1}{2}\right), N_d - N_2 + 1\right) \le i \le \min\left(N_d - N_1, \text{floor}\left(\dfrac{\beta}{2}\right)\right) \\ \max\left(\alpha - N_d, N_d - N_1 + 1\right) \le i \le \min\left(N_d - N_2, \beta - N_d + N_2\right) \\ \max\left(\alpha - N_d, N_d - N_1 + 1, N_d - N_2 + 1\right) \le i \le \text{floor}\left(\dfrac{\beta}{2}\right) \end{array} \right\}.$$

Finally, I_u is the union of 4 disjoint, possibly empty, intervals $I_1 \cup I_2 \cup I_3 \cup I_4$ where $I_k \cap I_l = \emptyset \ \forall \ k \neq l$ and:

$$I_1 = \left[\max\left(i_{\min}, \text{ceil}\left(\frac{\alpha - N_1}{2}\right)\right), \min\left(N_d - N_2, \beta - N_d + N_2, N_d - N_1, i_{\max}\right)\right],$$
$$I_2 = \left[\max\left(i_{\min}, \text{ceil}\left(\frac{\alpha - N_1}{2}\right), N_d - N_2 + 1\right), \min\left(N_d - N_1, \text{floor}\left(\frac{\beta}{2}\right), i_{\max}\right)\right],$$
$$I_3 = \left[\max\left(i_{\min}, \alpha - N_d, N_d - N_1 + 1\right), \min\left(N_d - N_2, \beta - N_d + N_2, i_{\max}\right)\right],$$
$$I_4 = \left[\max\left(i_{\min}, \alpha - N_d, N_d - N_1 + 1, N_d - N_2 + 1\right), \min\left(\text{floor}\left(\frac{\beta}{2}\right), i_{\max}\right)\right].$$

References

1 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. Nature Rev Genet 2005;6:95–108.
2 Zhao H: Family-based association studies. Stat Methods Med Res 2000;9:563–587.
3 Nielsen DM, Ehm MG, Weir BS: Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet 1999;63:1531–1540.
4 Armitage P: Tests for linear trends in proportions and frequencies. Biometrics 1955; 11:375–386.
5 Sasieni PD: From genotypes to genes : doubling the sample size. Biometrics 1997;53: 1253–1261.
6 Cochran WG: The chi-square test of goodness of fit. Annls Math Stat 1952;23:315–345.
7 Jackson MR, Genin E, Knapp M, Escary JL: Accurate power approximation for χ^2-tests in case-control association studies of complex disease genes. Annls Hum Genet 2002; 66:307–321.
8 Schaid DJ, Jacobsen SJ: Biased tests of association: Comparisons of allele frequencies when departing from Hardy-Weinberg proportions. Am J Epidemiol 1999;149:706–711.
9 Press WH, Teukolsky SA, Vetterling WT, Flannery BP: Numerical recipies in C. Error, Accuracy, and Stability, Cambridge University Press, 1992.
10 Matsuzaki H, Dong S, Loi H, Di X, Liu G: Genotyping over 100,000 snps on a pair of oligonucleotide arrays. Nature Methods 2004;1:109–111.
11 Hosking L, Lumsden S, Lewis K, Yeo, A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF: Detection of genotyping erros by Hardy-Weinberg 21 equilibrium testing. E J Hum Genet 2004;12:395–399.
12 Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289–300.
13 Longmate JA: Complexity and power in case-control association studies. Am J Hum Genet 2001;68:1229–1237.
14 Wright S: Systems of mating. Genetics 1921; 6:111–178.
15 Trikalinos TA, Salanti G, Khoury MJ, Ioannidis JPA: Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated genedisease associations. Am J Epidemiol 2006;163:300–309.
16 Storey JD, Tibshirani R: Statistical significane for genome-wide studies. PNAS 2003; 100:9440–9445.
17 Knapp M: Re: 'biased tests of association: comparisons of alleles frequencies when departing from Hardy-Weinberg proportions'. Am J Epidemiol 2003;153:287.
18 Page G, George V, Go R, Page P, Allison D: 'Are we there yet?': Deciding when one has demonstrated specific genetic causation in complex diseases and quantative traits. Am J Hum Genet 2003;73:711–719.
19 Shen H, Liu Y, Liu P, Recker RR, Deng HW: Nonreplication in genetic studies of complex diseases – lessons from studies of osteoporosis and tentative remedies. J Bone Mineral Res 2005;20:365–376.
20 Gordon D, Heath SC, Liu X, Ott J: A transmission/disequilibrium test that allows for genotyping errors in the analysis of single nucleotide polymorphism data. Am J Hum Genet 2001;69:371–380.
21 Devlin B, Roeder K: Genomic control for association studies. Biometrics 1999;55:997–1004.
22 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect populations stratificiations in association studies. Am J Hum Genet 1999;65:220–228.

Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 22

Detecting Local High-Scoring Segments: a First-Stage Approach for Genome-Wide Association Studies

Mickael Guedj* David Robelin† Mark Hoebeke‡
Marc Lamarine** Jérôme Wojcik†† Gregory Nuel‡‡

*Laboratoire Statistique et Génome, guedj@genopole.cnrs.fr
†Laboratoire Statistique et Génome, robelin@genopole.cnrs.fr
‡Laboratoire Statistique et Génome, hoebeke@genopole.cnrs.fr
**Serono Pharmaceutical Research Institute, marc.lamarine@serono.com
††Serono Pharmaceutical Research Institute, Jerome.Wojcik@serono.com
‡‡Laboratoire Statistique et Genome, nuel@genopole.cnrs.fr

Detecting Local High-Scoring Segments: a First-Stage Approach for Genome-Wide Association Studies*

Mickael Guedj, David Robelin, Mark Hoebeke, Marc Lamarine, Jérôme Wojcik, and Gregory Nuel

Abstract

Genetic epidemiology aims at identifying biological mechanisms responsible for human diseases. Genome-wide association studies, made possible by recent improvements in genotyping technologies, are now promisingly investigated. In these studies, common first-stage strategies focus on marginal effects but lead to multiple-testing and are unable to capture the possibly complex interplay between genetic factors.

We have adapted the use of the local score statistic, already successfully applied to analyse long molecular sequences. Via sum statistics, this method captures local and possible distant dependences between markers. Dedicated to genome-wide association studies, it is fast to compute, able to handle large datasets, circumvents the the multiple-testing problem and outlines a set of genomic regions (segments) for further analyses. Applied to simulated and real data, our approach outperforms classical Bonferroni and FDR corrections for multiple-testing. It is implemented in a software termed LHiSA for Local High-scoring Segments for Association and available at: http://stat.genopole.cnrs.fr/software/lhisa.

KEYWORDS: association studies, local score, sum statistics, SNP

---

*We would like to thank Serono for providing data as well as Bernard Prum's team for fruitful discussions; more particularly Franck Picard, Karl Forner and Hiroaki Tanaka.

Introduction

Localizing trait loci responsible for human diseases is a challenge for new drug target discovery. Limited results from linkage analysis scans and the increasing availability of dense Single Nucleotid Polymorphisms (SNP) maps (Collins et al. 1998) due to rapid improvements in SNP genotyping technologies (Shendure et al. 2004, Hirschhorn et al. 2005) have recently led geneticists towards genome-wide association studies with hopes of encouraging results concerning our understanding of the genetic basis of complex diseases (Risch and Merikangas 1996, Risch 2000, Carlson et al. 2004, Hirschhorn et al. 2005). Association analyses can be carried out in a family-based framework using the Transmission Disequilibrium Test (TDT) (Spielman and Ewens 1996) for instance, or in a case-control design where individuals are unrelated, sampled from the same general population, and separated into cases and controls. The TDT method has been shown to be more robust to confounding factors such as stratification but also more susceptible to technological false-positive results (Ott 2004). Gordon et al. (2001) have proposed an approach to tackle genotyping errors in a TDT analysis. However case-control studies are still largely preferred mainly because they require less cost and time to collect data (McGinnis et al. 2002). They merely rely on the fact that genotypes associated with the disease should accumulate among cases compared to controls.

Nevertheless, association analyses appear to be insufficient to shed light on the causative genes of non-mendelian diseases. One reason pointed out by Page et al. (2003) is that complex diseases result from the interplay of multiple genetic and environmental factors. As a result each causal gene only makes a small contribution to overall heritability and "disease" alleles may be present in the healthy population. Classical marker-by-marker analyses based on contingency tables can then hardly succeed, since they rely on the detection of marginal effects. In addition, the statistical problem of multiple-testing arises, inflating the number of false positive results (type-I errors) when a large number of SNPs are tested independently. To face such complexity, statisticians and computer scientists have developed numbers of multi-marker strategies that can be divided into two classes, as underlined by Hoh and Ott (2003): "multipoint analysis" for the joint analysis of neighbouring marker loci, with the purpose of localizing one disease locus independently from the others (Morris et al. 2003, Tregouet et al. 2004, Schaid 2004) and "multi-locus" approaches that are specifically designed to find multiple disease loci, possibly on different chromosomes (Nelson et al. 2001, North et al. 2003, Yoon et al. 2003, Bastone et al. 2004, Bureau et al. 2005). Such strategies appear very promising in considering various interaction patterns between loci, but present

some weaknesses (Wille et al. 2003): First, most of them are not appropriate to treat data of large dimensionality. Moreover these methods require generally a difficult and unintuitive specification of parameters. Finally the statistical assessment of the results is not always available. See Hoh and Ott's review (2003) on mathematical multi-locus approaches for more detailed discussions.

Consequently, further efforts have to be made to propose methods that are easy to use, computationally fast, able to handle large data sets while reducing the multiple-testing problem. Sum statistics based methods can help to meet these needs. Based on simulation studies conclusions, Longmate (2001) suggested that combining single-marker genetic marginal effects should increase the power to detect susceptibility genes, possibly interacting with each other. One approach to combine individual association scores without assuming any interaction pattern among markers can be done with sums. Two examples of such strategies termed "scan statistics" (Hoh and Ott 2000) and "set association" (Hoh et al. 2001) have been proposed: they have been shown to be more powerful than conventional single-point analyses and presented promising results when applied to real data (Whille et al. 2003, Hao et al. 2004).

As part of the sum statistics family methods for association, we propose the use of the local score. This statistic has been a matter of theorical works (Altschul and Erickson 1986a, Altschul and Erickson 1986b, Karlin and Altschul 1990, Dembo and Karlin 1991, Karlin and Dembo 1992, Mercier and Daudin 2001) and frequently applied to long DNA or protein sequences to locate transmembrane or hydrophobic segments, DNA-binding domains, regions of concentrated charges (Karlin et al. 1991, Brendel et al. 1992, Karlin and Brendel 1992). It has been extended to identify similar regions among two or more sequences (Altschul et al. 1990) . See Karlin (2005) for a review on the topic.

In association studies, the local score approach intends to delimit segments (or subsequences) presenting unexpected accumulations of high scores, assuming that such features may be biologically relevant and outline etiological genomic regions. It represents a natural improvement of sliding-frame methods since it does not require to specify the size of the segments. Via sum statistics, we will see how this method considers local information to identify each segment and can be thus termed as a multi-point and multi-locus strategy. It proposes at the end a selection of candidate regions through the genome associated with a measure of statistical significance.

Methods

Definition

Let $\mathbb{X} = (X_i)_{i=1,...,n}$ a sequence of real random variables:

$$H = \max_{1 \leqslant i \leqslant j \leqslant n} \left(\sum_{i}^{j} X_k \right)$$

defines the local score assigned to $\mathbb{X}$ (Figure 1). In practice, the local score statistic is defined as the value of the segment with the maximal sum of scores X_i; this segment is also termed as "maximal scoring subsequence", "locally optimal subsequence" or "maximal sum interval" in the literature. Note that the segment is maximal in the sense that it can not be extended or shortened without reducing the local score. Consequently the variables X_i must have a negative expectation otherwise the maximal segment would easily span the entire sequence, which is not consistent with the aim of this approach.

This definition of the local score restrains our search to the highest scoring sequence but data may not contain only one region of interest: next highest scoring subsequences are as potentially interesting. However, segments with the successive high scores could overlap the highest-scoring one and lead to the identification of more or less the same genomic region, yielding to no additional useful information. Consequently, it appears better to look for disjoint segments, with the k^{th} best segment defined as the local score of the initial sequence disjoint from the preceding $k-1$ best segments. We consider in this case $H^{(1)} \geqslant ... \geqslant H^{(k)}$ as being the scores of the k first and distinct highest-scoring segments.

Test for association

Given the local score statistic, advantages to simple-marker strategies will arise from the definition of a more appropriate alternative hypothesis (H_1): there is at least one contiguous segment of the genome associated with the disease. This alternative is justified by haplotypes that have remained intact for many generations around a disease susceptibility locus (DSL). As a result, instead of looking for one marker at a time, it has been shown relevant to look for a stretch of DNA inherited in all mutation carriers (Clayton et al. 2004). From this point of view, our approach can be categorized as a haplotype similarity based method (Tzeng et al. 2003). Note that when more than one mutation is suspected to alter the gene function, we expect an accumulation of high scores of association in a quite small genomic region, situation also well detected by the local score statistic.

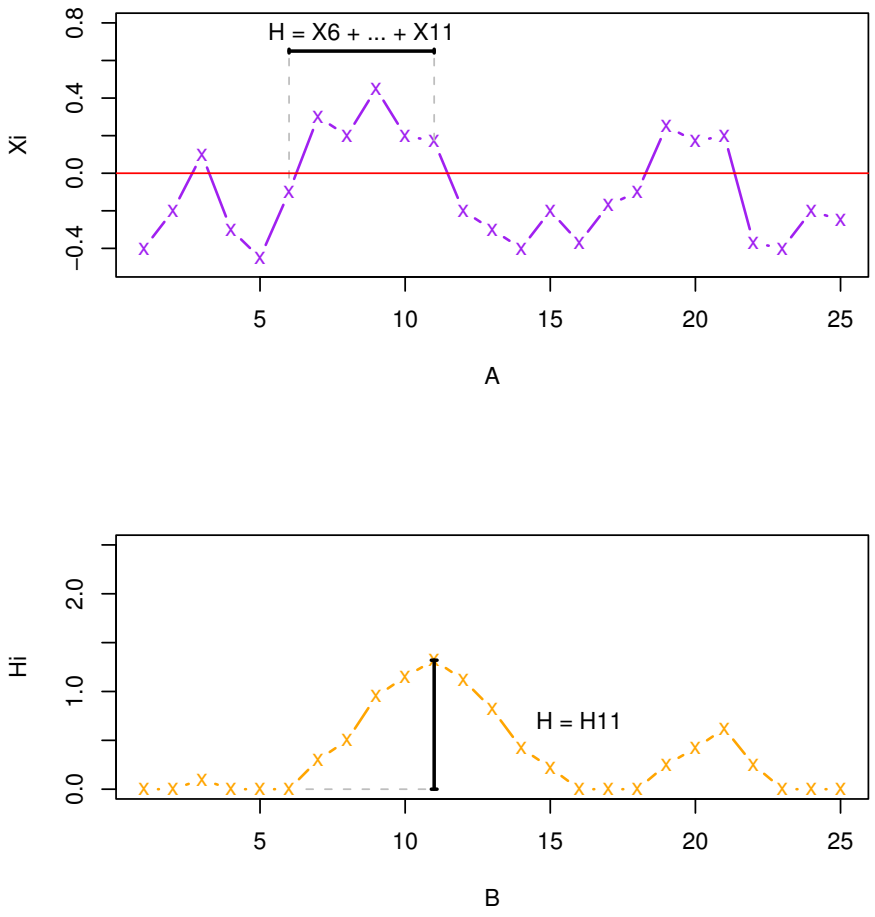


Figure 1. Local score statistic : A - A set of scores X_i represents the sequence $\mathbb{X}$. The highest-scoring segment is delimited. A naive approach to get the local H score from $\mathbb{X}$ consists in comparing the value of $\sum_i^j X_k$ for all possible pairs $(i,j)_{i \leqslant j}$; such an approach is quadratic with the length of the sequence. **B -** Considering the processus $\mathbb{H} = (H_i)_{1,\dots,n}$ with $H_i = \max(0, H_{i-1} + X_i)$ and $H_0 = \max(0, X_0)$, the local score statistic is merely defined as $H = \max(H_i)$. Finding the maximal scoring subsequence comes down to find the maximal value $\mathbb{H}$ what is linear instead of quadratic.

Statistical results

Based on the extreme values theory and the work of Iglehart (1972) , Karlin and Dembo (1992) underlined that when X_i are independent, identically distributed and $\mathbb{E}(X_i) < 0$, the distribution of H can be well approximated for large n by the Gumbel distribution:

$$\lim_{n\to+\infty} \mathbb{P}\left(H \geqslant \frac{\log n}{\lambda} + x\right) = 1 - \exp\left(-Ke^{-\lambda x}\right) \tag{1}$$

The random distribution of the local score is thus expressed in terms of three parameters that are the length of the sequence n and two parameters of normalization, K and λ depending on the distribution of scores X_i. Given all parameters, the statistical theory can be more simply expressed by the use of the normalized score $H' = \lambda H - \log(nK)$:

$$\lim_{n\to+\infty} \mathbb{P}\left(H' \geqslant x\right) = 1 - \exp\left(-e^{-x}\right) \tag{2}$$

When searching for multiple genomic regions, $H^{(1)}, ..., H^{(k)}$ can also be written using normalized expressions $H'^{(1)}, ..., H'^{(k)}$ with $H'^{(i)} = \lambda H^{(i)} - \log(nK)$ for which Karlin and Altschul (1993) and Altschul (1997) provided some interesting results.

Considering values $h'_1 \geqslant h'_2 \geqslant ... \geqslant h'_k$, the asymptotic joint probability density function for $H'^{(1)} \geqslant ... \geqslant H'^{(k)}$ is well approximated by:

$$\lim_{n\to+\infty} f(h_1, ..., h_k) = \exp\left(-e^{-h_k} - \sum_{i=1}^{k} h_i\right) \tag{3}$$

From this, Karlin and Altschul (1993) deduced a generalization of Formula 1 for the k^{th} score:

$$\lim_{n\to+\infty} \mathbb{P}\left(H'^{(k)} \geqslant h'_k\right) = 1 - \exp\left(e^{-x}\right) \times \sum_{i=0}^{k-1} \frac{e^{-ix}}{i!} \tag{4}$$

The authors also provided the asymptotic probability density function for the sum of the k highest normalized scores $T'^{(k)} = H'^{(1)} + ... + H'^{(k)}$:

$$\lim_{n\to+\infty} f_k(t') = \frac{e^{-t'}}{k!(k-2)!} \int_0^{+\infty} y^{k-2} \exp\left(-e^{(y-t')/k}\right) dy \tag{5}$$

Algorithm

Knowledge about the local score is compiled in the following four-step algorithm (Figure 2) in order to propose a new approach for SNP-data analysis:

1 - First we assign an "individual score" X_i to each marker, a high score meaning a high chance of association to the disease. It can be based on common statistics like allelic, genotypic, Hardy-Weinberg or Armitage statistics; p-values p_i can as well be used but need a transformation such as $X_i = -\log_{10}(p_i)$. As previously underlined, a constraint of the strategy is to have negative expected individual scores; that does not happen with scores based on chi-square statistics or p-values for instance. In such cases, a constant δ must be subtracted from the whole signal $\mathbb{X}$ to get a corrected sequence $\mathbb{X}' = (X'_i)_{i=1,...,n}$, with $X'_i = X_i - \delta$ and $\mathbb{E}(X'_i) < 0$. SNPs with a score higher than δ will improve the cumulative score of a given segment whereas SNPs with a score below the threshold will penalize it.

2 - Then the aim is to identify the best high-scoring segments and to compute their respective scores $H^{(1)}, ..., H^{(k)}$. A simple strategy is to use an iterative algorithm: **(i)** find the highest-scoring segment (refer to Figure 1), **(ii)** remove it, **(iii)** apply the algorithm again. The process is stopped when the next best score is non-positive. In the worst case, all maximal scoring subsequences can be found in $O(n^2)$ time. Considering the length of the sequences we want to analyze, we will prefer instead of this naive approach, a more recent and faster one developed by Ruzzo and Tompa (1999) that finds all maximal scoring subsequences in a linear running time, and which has been proved to be 15 to 20 times faster on megabase sequences. From this step we get cumulative scores $H^{(1)}, ..., H^{(k)}$. Testing n markers has been reduced to the test of k sums.

3 - Given $H^{(1)}$, ...,$H^{(k)}$, the third step proposes a way to select a set of interesting segments. The successive local scores are combined into new sum statistics $T^{(1)}$, ..., $T^{(k)}$ with $T^{(i)} = H^{(1)} + ... + H^{(i)}$. Corresponding p-values $p_{T1}, ..., p_{Tk}$ can be computed using theoretical results and normalized scores (we obtain the tail probability that $T'^{(k)} \geqslant x$ integrating formula 5 for t' from x to infinity) or by Monte-Carlo simulations (permuting case and control labels). As proposed by Karlin and Altschul (1993), we select interesting segments as being the r first ones with $r = \text{argmin}_i(p_{Ti+1} > p_{Ti})$. In practice, it makes the hypothesis that segments are biologically interesting as long as they improve the overall result (p_{Ti} decreases) and segments that are actually noise do not improve it anymore. This gives us the number r of segments to select and an associated statistic $p^{(0)}_{\min} = p_{Tr}$ for this selection.

4 - The last step consists of assessing the global significance (p_G) of the

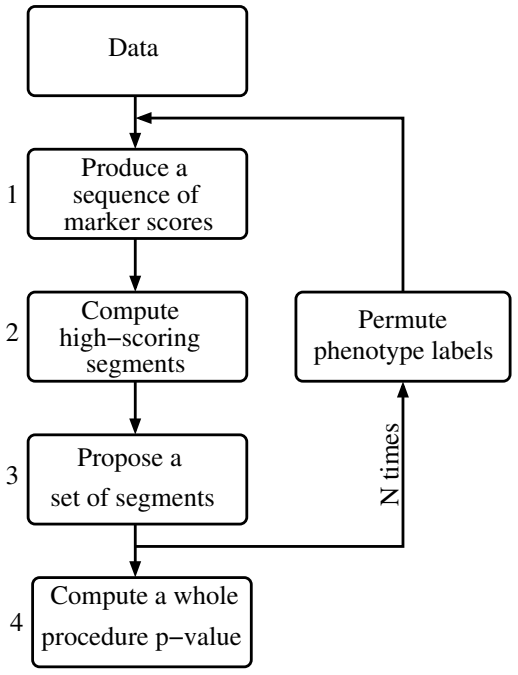


Figure 2. Algorithm

process via Monte-Carlo simulations. We iterate N times steps 1 to 3, permuting each time case and control labels, and computing $p_{\min}^{(i)}$ attached to the i^{th} iteration. This operation leads to the distribution of $p_{\min}$ under the assumption H_0 of total independence between genotypes and the disease. Finally, the p-value of the whole procedure is:

$$p_G = \frac{\text{card}\left\{i, p_{\min}^{(i)} \leqslant p_{\min}^{(0)}\right\}}{N}$$

Our method circumvents the multiple-testing problem by reducing the number of tests performed from n marker scores (X_i) to k local scores $(H^{(i)})$ and then from k to one statistic $(p_{\min}^{(0)})$.

Power study

This power study aims at showing the advantages of such a strategy compared to classical simple-marker with Bonferroni and FDR (Benjamini and Hochberg 1995) corrections. In addition we investigated the effect of the parameter δ

on results, as well as the efficiency of our selection when more than one DSL contributes to the phenotype.

Computing power

The power calculations were performed using Monte-Carlo simulations based on real data by applying a genetic model to the functional site(s). We used diplotype frequencies of a given population as an empirical distribution of the range of possible diplotypes. This way of simulating preserves the LD patterns expected in real data (Chapman et al. 2003, Nielsen et al. 2004). We then use the following genetic model to generate samples of affected and unaffected individuals, along with their genotypes at each marker of the dataset. Let a bi-allelic disease susceptibility locus with its two possible alleles (the allele of susceptibility A and a), p be the frequency of the allele A and r_0, r_1 and r_2 the frequencies of the genotypes aa, aA and AA respectively. According to Wright's model (1921), we have: $r_2 = p^2 + fp(1-p)$, $r_1 = 2p(1-p) - 2fp(1-p)$ and $r_0 = (1-p)^2 + fp(1-p)$ with f the coefficient of consanguinity. Now we introduce the prevalence of the disease (K_p), and the penetrances (f_i) associated to each genotype (i). Considering the relative risks (RR_i) such as $RR_i = \frac{f_i}{f_0}$ for $i = 1$ or 2, we define the four main modes of inheritance (MOI) underlying the mode of action of the DSL on the disease: recessive ($RR_1 = 1$), multiplicative ($RR_1 = \sqrt{RR_2}$), additive ($RR_1 = \frac{RR_2+1}{2}$) and dominant ($RR_1 = RR_2$). From these parameters, we can easily derive $f_0 = K_p/(r_0 + RR_1.r_1 + RR_2.r_2)$, $f_1 = RR_1.f_0$ and $f_1 = RR_1.f_0$ and finally genotype frequencies for the DSL in case and control populations:

$$\begin{aligned}(p_{d_0}, p_{d_1}, p_{d_2}) &= \left(\frac{f_0.r_0}{K_p}, \frac{f_1.r_1}{K_p}, \frac{f_2.r_2}{K_p}\right)\\ (p_{c_0}, p_{c_1}, p_{c_2}) &= \left(\frac{(1-f_0).r_0}{1-K_p}, \frac{(1-f_1).r_1}{1-K_p}, \frac{(1-f_2).r_2}{1-K_p}\right)\end{aligned}$$

Simulations

Simulations were based on 674 SNPs covering the chromosome 19 and DNA from 301 controls were genotyped using the 100K Affymetrix chip. DSL were chosen randomly as long as they satisfy polymorphim constraints (minor genotype frequency more than 5%) and random union of gametes was assumed ($f = 0$). We computed the genotypic Pearson statistic for each SNP and compared the local score strategy with common Bonferroni and FDR corrections. The sampling and testing procedure was repeated 200 times for an additive MOI, different values of RR_2 and δ, first with one (hidden or not) DSL and then with three.

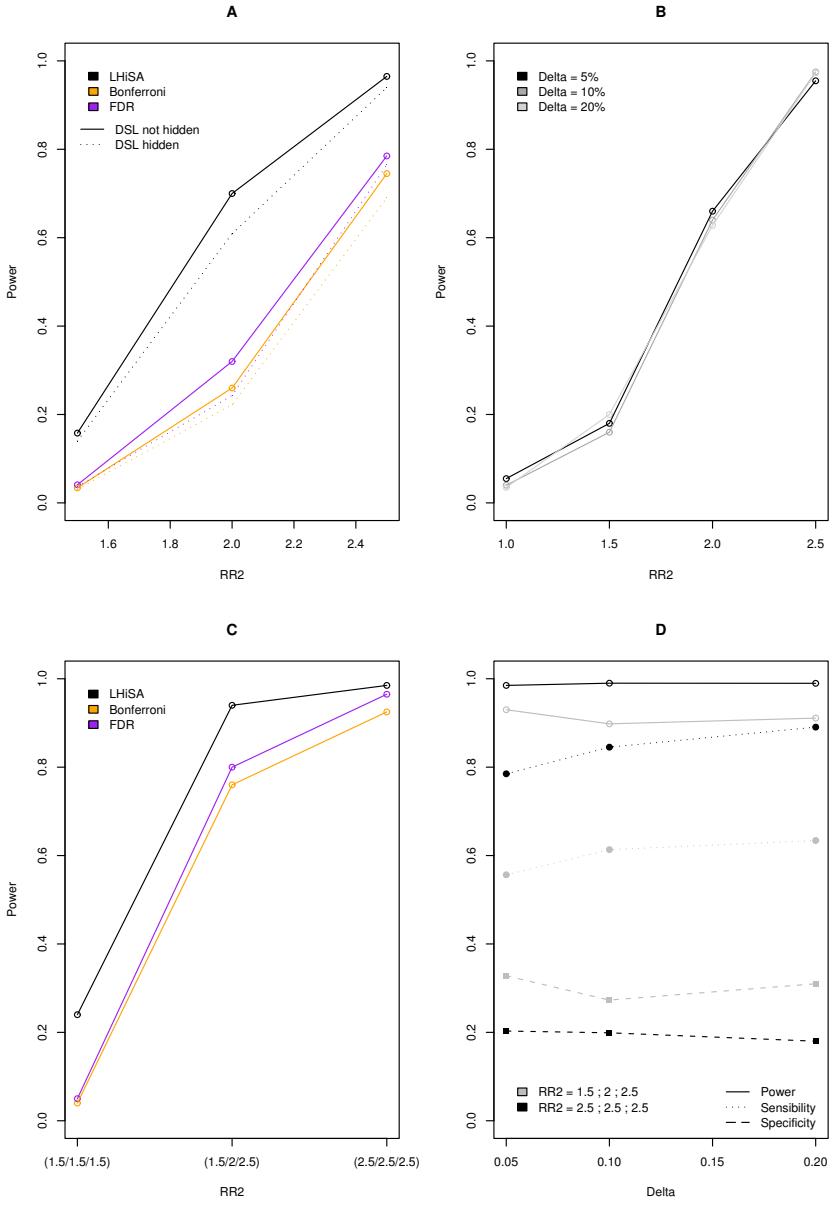


Figure 3. Power study: A - LHiSA, Bonferroni and FDR power estimates for three H_1 ($RR_2 = 1.5$, 2 and 2.5), DSL hidden or not. **B -** LHiSA power estimates for H_0 ($RR_2 = 1$) and three H_1 ($RR_2 = 1.5$, 2 and 2.5) according to δ. **C -** Comparison of power for three DSL H_1 (RR_2 =(1.5/1.5/1.5), (1.5/2/2.5) and (2.5/2.5/2.5)). **D -** LHiSA power, specificity and sensibility estimates for two H_1 ($RR_2 = (1.5/2/2.5)$ and $(2.5/2.5/2.5)$) according to δ.

Results

Results are summarized in Figure 3. In the two situations tested (DSL hidden or not, Figure 3-A), our strategy shows better performances than the classical Bonferroni and FDR ones with an average increase of 0.25. Simulating a hidden DSL logically leads to a loss of power. Note also that FDR is more powerful than Bonferroni which is consistent with the literature. On situations considered, the parameter δ does not seem to have any effect on the power (Figure 3-B). Moreover, under H_0 ($RR_2 = 1$), the type-I error-rate remains around 5% for any value of δ which is reassuring concerning the statistical assessment of results. When dealing with three DSL, the local score strategy also presents greater power (Figure 3-C). In this case, δ does not seem to have more effect on the probability to reject H_0 (Figure 3-D). However, the ability of the method to detect the true positives (sensitivity) increases and the count of true positives among the positives (specificity) tends to decrease along with δ.

Application

Data

We applied our strategy to genetic data implicating G72 and DAAO (D-amino acid oxydase) genes in schizophrenia (Chumakov et al. 2002). The dataset consists of 172 SNPs: 8 on chromosome 12 (surrounding DAAO) and 164 on chromosome 13 (including G72). DNA from 213 schizophrenic patients and 241 normal individuals were genotyped with this marker set. We computed the allelic chi-square statistics for each SNP (i) and associated p-values (p_i). Let $X_i = -\log_{10}(p_i)$ and $X'_i = X_i - \delta$ with $\delta = -\log_{10}(0.1)$ so that we consider SNPs with a p-value lower than 0.1 as potentially interesting and contributing positively to the score of a segment (Figure 4). The implication of G72 in schizophrenia has been reported many times in the literature (Detera-Wadleigh and McMahon, 2006)

Results

The local score strategy identifies three segments ($r = 3$), and $p_{\min}^{(0)} = 0.1660$ corresponding to the sum of the three first segments (Table 1). The global significance is $p_G = 0.22$. The two first segments are part of the G72 gene and the third one is part of the DAAO gene. Our method success in detecting the two genes expected. However, the significance is not convaincing enough to conclude to an association and confirms recent doubts about the implication of G72 and DAAO in schizophrenia (Riley and Kendler 2006).

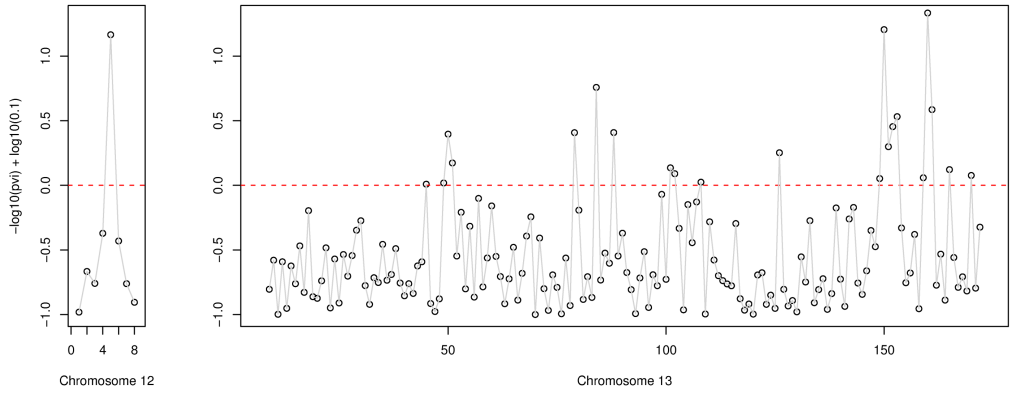


Figure 4. Schizophrenia - data: for each SNP i, an individual score $X_i = -log_{10}(p_i) + log_{10}(0.1)$ is assigned. SNPs labelled 1 to 8 are located on chromosome 12 and SNPs labelled 9 to 172 on chromosome 13.

rank	chr	segment	H	T	p_T
1	13	149-153	2.542	2.542	0.2459
2	13	159-161	1.978	4.520	0.1737
3	12	5	1.165	5.686	0.1660
4	13	84	0.758	6.444	0.1702
5	13	49-51	0.587	7.031	0.1747

Table 1
Schizophrenia - results of the local score strategy: $N = 10000$, $\delta = -\log_{10}(0.1)$, $p_{min}^{(0)} = 0.1660$ *and* $p_G = 0.22$.

Discussion

Based on $H^{(i)}$ and $T^{(i)}$, the sums of contiguous marker scores and the combination of distant segments, our strategy proposes to test the null hypothesis of no association against the existence of at least one region involved in the disease. The size of the regions is not predefined. This is an important advantage over sliding-frame strategies (Hoh and Ott 2000, Marques-Bonet et al. 2005) for which the choice of the frame size is arbitrary, often difficult, and may depend on multiple biological *a priori* (pattern of LD and density of markers for instance). Additionally, there is absolutely no biological reason to constrain the successive high-scoring segments to the same size. Our method handles the multiple-testing problem by reducing the number of tests from n markers to one statistic $p^{(0)}_{\min}$. On our power study it outperforms classical Bonferroni and FDR strategies. On schizophrenia data, it detects expected genes but with a significance that does not allow us to really conclude to an association with the disease.

The method requires the specification of one control parameter δ. This parameter represents the level upon which one marker is considered to be putatively interesting and hence can be intuitively set. One can choose classical 1% or 5% levels for the considerate marker score (X_i). However, since effects we want to detect are generally weak, we decide to increase it to 10%. Results predictively change along with this parameter: lowering δ will logically increase the size of the segments and the number of segments detected whereas augmenting it will lead to the contrary. Moreover, to correctly determinate the significance of the process using the Gumbel approximation, it is a necessary assumption for marker scores X_i to be weakly dependent. Because of the LD that may exist between successive loci (particulary for dense maps), this assumption is not ensured, so we recommend for this step to use Monte-Carlo simulations instead of the extreme value theory. It also has the advantage of avoiding asymptotic approximations and hence works on small sequences. The software has been implemented so that the use of Monte-Carlo simulations does not dramatically extend the algorithm computation time. The theoretical time complexity is in $O\left(N\log(N) + Nn\log(n)\right)$ and the memory complexity is in $O\left(n + N\log(n)\right)$. On a quite basic machine (Intel Pentium 4 CPU 2.80GHz, 512Mo RAM) it takes approximatively 10, 180 and 800 seconds to process respectively 200, 2000 and 10000 SNPs with $N = 2000$ simulations. The use of a more powerful machine will obviously improve this computation time, and considering the large number of independent Monte-Carlo simulations performed, our algorithm allows for parallelisation if applied on a cluster of machines. Finally, our method handles data from SNP genotyping but can

easily be adapted to other frameworks and types of markers.

Developed to manage large data sets in a quite reasonable time and to propose a selection of segments surrounding putative etiological sites, we believe that our approach is well adapted to genome-wide association studies. Having found significant associations with given regions, this selection can then be subject to further attention (haplotype reconstruction, fine-mapping, annotation, interaction patterns...) in order to formulate assumptions on the way they may contribute to the disease.

References

Altschul, S. (1997). *Theorical and compuational methods in genome research*, chapter Evaluating the statistical significance of multiple distinct local alignments, pages 1–14. New York: Plenum Press.

Altschul, S. and Erickson, B. (1986a). Locally optimal subalignments similarity and its significance levels. *Bulletin of Mathematical Biology* **48**, 633–660.

Altschul, S. and Erickson, B. (1986b). A nonlinear measure of subalignments similarity and its significance levels. *Bulletin of Mathematical Biology* **48**, 617–632.

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990). Basic local alignement search tool. *Journal of Molecular Biology* **215**, 403–410.

Bastone, L., Reilly, M., Rader, D. and Foulkes, A. (2004). Mdr and prp: a comparison of methods for high-order genotype-phenotype associations. *Human Heredity* **58**, 82–92.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300.

Brendel, V., Bucher, P., Nourbakhsh, I., Blaisdell, B. and Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *Proceedings of the National Academy of Science USA* **89**, 2002–2006.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K., Hayward, B., Keith, T. and Van Eerdewegh, P. (2005). Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology* **28**, 171–182.

Carlson, C., Eberle, M., Kruglyak, L. and Nickerson, D. (2004). Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452.

Chapman, J. M., Cooper, J. D., Todd, J. A. and Clayton, D. G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human Heredity* **56**, 18–31.

Chumakov, I., Macciardi, F., Sham, P., Straub, R., Weinberger, D., Cohen, N. and Cohen, D. (2002). Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. *Proceedings of the National Academy of Science USA* **99**, 13675–13680.

Clayton, D., Champan, J. and Cooper, J. (2004). Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology* **27**, 415–428.

Collins, F., Brooks, L. and Chakravarti, A. (1998). A dna polymorphism discovery resource for research on human genetic variation. *Genome Research* **8**, 1229–1231.

Dembo, A. and Karlin, S. (1991). Strong limit theorems of empirical functionals for large exceedances of partial sums of iid variables. *Annals of Probability* **19**, 1737–1755.

Detera-Wadleigh, S. D. and McMahon, F. J. (2006). G72/g30 in schizophrenia and bipolar disorder: review and meta-analysis. *Biological Psychiatry [in press]* .

Gordon, D., Heath, S., Liu, X. and Ott, J. (2001). A transmission/disequilibrium test that allows for genotyping errors in the analysis of single nucleotide polymorphism data. *American Journal Human Genetics* **69**, 371–380.

Hao, K., Xu, X., Laird, N., Wang, X. and Xu, X. (2004). Power estimation of multiple snp association test of case-control study and application. *Genetic Epidemiology* **26**, 22–30.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95–108.

Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes. *Proceedings of the National Academy of Science USA* **97**, 9615–9617.

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics* **4**, 701–709.

Hoh, J., Wille, A. and Ott, J. (2001). Trimming, weighting, and grouping snps in human case-control association studies. *Genome Research* **11**, 2115–2119.

Iglehart, D. (1972). Extremes values in the gi/g/1 queues. *Annals of Mathematical Statistics* **43**, 627–635.

Karlin, S. (2005). Statistical signals in bioinformatics. *Proceedings of the National Academy of Science USA* **102**, 13355–13362.

Karlin, S. and Altschul, S. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Science USA* **87**, 2264–2268.

Karlin, S. and Altschul, S. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Science USA* **90**, 5873–5877.

Karlin, S. and Brendel, V. (1992). Chance and significance in protein and dna sequence analysis. *Science* **257**, 39–49.

Karlin, S., Bucher, P., Brendel, V. and Altschul, S. (1991). Statistical-methods and insights for protein and dna-sequences. *Annual Review of Biophysics and Biophysical Chemistry* **20**, 175–203.

Karlin, S. and Dembo, A. (1992). Limit distributions of maximal segmental score among markov-dependant partial sums. *Advances in Applied Probability* **24**, 113–140.

Longmate, J. (2001). Complexity and power in case-control association studies. *American Journal Human Genetics* **68**, 1229–1237.

Marques-Bonet, T., Lao, O., Goertsches, Robert ad Comabella, M., Montalban, X. and Navarro, A. (2005). Association cluster detector: a tool for heuristic detection of significance cluters in whole-genome scans. *Bioinformatics* **21**, ii180–ii181.

McGinnis, R., Shifman, S. and Darvasi, A. (2002). Power and efficiency of the tdt and case-control design for association scans. *Behavior Genetics* **32**, 135–144.

Mercier, S. and Daudin, J. (2001). Exact distribution for the local score of one i.i.d random sequence. *Journal of Computational Biology* **8**, 373–380.

Morris, A., Whittaker, J. and Balding, D. (2003). Multipoint linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity. *Proceedings of the National Academy of Science USA* **11**, 13442–13446.

Nelson, M., Kardia, S., Ferrell, R. and Sing, C. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research* **11**, 458–470.

Nielsen, D. M., Ehm, M. G., Zaykin, D. V. and Weir, B. S. (2004). Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics* **168**, 1029–1040.

North, B., Curtis, D., Cassell, P., Hitman, G. and Sham, P. (2003). Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes. *Annals Human Genetics* **67**, 348–356.

Ott, J. (2004). Issues in association analysis: error control in case-control association studies for disease gene discovery. *Human Heredity* **58**, 171–

174.

Page, G., George, V., Go, R., Page, P. and Allison, D. (2003). "are we there yet?": deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *American Journal Human Genetics* **73**, 711–719.

Riley, B. and Kendler, K. (2006). Molecular genetic studies of schizophrenia. *European Journal of Human Genetics* **14**, 669–680.

Risch, N. (2000). Searching for genes in complex diseases: lessons from systemic lupus erythematosus. *American Society for Clinical Investigation* **105**, 1503–1506.

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

Ruzzo, W. and Tompa, M. (1999). A linear time algorithm for finding all maximal scoring subsequences. In *7th International Conference of Intelligent Systems for Molecular Biology*, pages 234–241.

Schaid, D. (2004). Evaluating association of haplotypes with traits. *Genetic Epidemiology* **27**, 348–364.

Shendure, J., Mitra, R., Varma, C. and Church, G. (2004). Advanced sequencing technologies: methods and goals. *Nature Review Genetics* **5**, 335–344.

Spielman, R. and Ewens, W. (1996). The tdt and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics* **59**, 983–989.

Tregouet, D., Escolano, S., Tiret, L., Mallet, A. and Golmard, J. (2004). A new algorithm for haplotype-based association analysis: the stochastic-em algorithm. *Annals of Human Genetics* **68**, 165–177.

Tzeng, J.-Y., Devlin, B., Wasserman, L. and Roeder, K. (2003). On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *American Journal of Human Genetics* **72**, 891–902.

Wille, A., Hoh, J. and Ott, J. (2003). Sum statistics for the joint detection of multiple disease loci in case-control association studies with snp markers. *Genetic Epidemiology* **25**, 350–359.

Wright, S. (1921). Systems of mating. *Genetics* **6**, 111–178.

Yoon, Y., Song, J., Hong, S. and Kim, J. (2003). Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. *Clinical Chemestry and Laboratory Medecine* **41**, 529–534.

doi: 10.1111/j.1469-1809.2006.00316.x

Computing power in case-control association studies through the use of quadratic approximations: application to meta-statistics

M. Guedj,[1,2]* E. Della-Chiesa,[1] F. Picard[1] and G. Nuel[1]
[1]*Laboratoire Statistique et Genome, FRANCE*
[2]*Serono, FRANCE*

Summary

In the framework of case-control studies many different test statistics are available to measure the association of a marker with a given disease. Nevertheless, choosing one particular statistic can lead to very different conclusions. In the absence of a consensus for this choice, a tempting option is to evaluate the power of these different statistics *prior* to make any decision. We review the available methods dedicated to power computation and assess their respective reliability in treating a wide range of tests on a wide range of alternative models.

Considering Monte-Carlo, non-central chi-square and Delta-Method estimates, we evaluate empirical, asymptotic and numerical approaches. Additionally we introduce the use of the Delta-Method, extended to order 2, intended to provide better results than the traditional order-1 Delta-Method. Supplementary data can be found at: http://stat.genopole.cnrs.fr/software/dm2.

Keywords: Power, association tests, Delta-Method

Introduction

Case-control association studies are considered to be the simplest framework to help elucidate the genetic basis of complex diseases (Risch, 2000). Even if they have some weaknesses with ragard to potential confounding factors such as population stratification, they remain an important tool in genetical epidemiology and are often preferred to family-based studies (Zhao, 2000) due to the availability of data. Such an approach involves unrelated individuals split into cases, who are diagnosed with the disease of interest, and unaffected controls. This merely relies on the assumption that disease-related genetic determinants should accumulate among cases.

Tests of association are used as a first step in the analysis process. Various tests are proposed based on either genotypes (Table 1) such as the genotypic, Hardy-Weinberg equilibrium or Cochran-Armitage tests, or alleles such as the allelic test.

Since different single statistics can be used to test for association, another strategy is to combine them via meta-statistics with the hope of gaining power.

If they aim to establish an association between markers and disease, each test has a slightly different null hypothesis (H_0) and hence a different efficiency with respect to the underlying hypothesis. One way to compare them is to assess their power (π), defined as the ability of a test to reject the null hypothesis when the alternative hypothesis (H_1) is true. Power studies require the distribution of statistics under H_1.

This article reviews and discusses the most commonly used mathematical frameworks to approximate the H_1 distribution (and hence to compute π) in the context of genetic association studies. Our study includes two popular approaches. The first is empirical and based on Monte-Carlo simulations under the alternative hypothesis. The second is based on the

*Corresponding author: Mickaël Guedj, Laboratoire Statistique et Genome, 523 place des terrasses de l'Agora, 91000 EVRY, Tel: +33-1-6087-3800, Fax: +33-1-6087-3809, E-mail: guedj@genopole.cnrs.fr

Table 1 The genotypic contingency table

	aa	*aA*	*AA*	total
Diseased	D_0	D_1	D_2	n_D
Control	C_0	C_1	C_2	n_C
Total	n_0	n_1	n_2	n

asymptotic non-central chi-square distribution of the statistics under H_1. We compare these two approaches with the Delta-method with emphasis on its extension to order 2. As expected, non-central chi-square approximations appear to be very reliable (whenever available), while the order-1 Delta-Method is not. To treat non-explicit cases (combinations of statistics, for instance) for which computationally expensive Monte-Carlo simulations are usually considered, we show that the order-2 Delta-Method approximations are sufficiently efficient to represent a valid and cheaper alternative.

Method

Testing for association

Let us denote by x a case-control sample that is a realisation of the random variable X, and which can be represented by a genotypic contingency table (Table 1). To establish an association we consider a null hypothesis (H_0) used to test a particular distribution of the observations. To do so we consider a statistic defined as a function of these observations: $\mathcal{S} = f(X)$, and carefully chosen such that $\mathcal{S}$ grows when H_0 is less likely. Using the distribution of $\mathcal{S}$ under H_0, we can find a threshold (t_α) such as $\alpha = \mathbb{P}_{H_0}(\mathcal{S} \geq t_\alpha)$, where level α can be set to 5% for example.

Computing power

Using the distribution of $\mathcal{S}$ under an alternative association hypothesis H_1, we can compute the power $\pi(\alpha)$ of the test such that: $\pi(\alpha) = \mathbb{P}_{H_1}(\mathcal{S} \geq t_\alpha)$. To calculate $\pi(\alpha)$, the first step is to define a genetic model as well as the null and alternative hypotheses.

Genetic Model

Consider a bi-allelic disease susceptibility locus (DSL) with A being the susceptibility allele and a the other, p the frequency of allele A, and r_0, r_1 and r_2 the genotype frequencies in the general population. Assuming that the Hardy-Weinberg equilibrium (HWE) holds in the population, genotypic frequencies reduce to $r_2 = p^2$, $r_1 = 2p(1-p)$ and $r_0 = (1-p)^2$. Now we introduce the prevalence of the disease (K_p), and penetrances (f_i) associated with each genotype (i). Considering the relative risks (RR_i) such that $RR_i = \frac{f_i}{f_0}$ for $i = 1$ or 2, we define the four main modes of inheritance (MOI) corresponding to the modes of action of the DSL on the disease: recessive ($RR_1 = 1$), multiplicative ($RR_1 = \sqrt{RR_2}$), additive ($RR_1 = \frac{RR_2+1}{2}$) and dominant ($RR_1 = RR_2$).

Considering these parameters, we can easily derive $f_0 = K_p/(r_0 + RR_1.r_1 + RR_2.r_2)$ and $f_i = RR_i.f_0$ for $i = 1$ or 2. With the further assumption of an infinite population, the genotype distributions (D_0, D_1, D_2) in cases and (C_0, C_1, C_2) in controls are multinomial with parameters:

$$(D_0, D_1, D_2) \sim \mathcal{M}\left(n_D; \frac{f_0 r_0}{K_p}, \frac{f_1 r_1}{K_p}, \frac{f_2 r_2}{K_p}\right),$$

$$(C_0, C_1, C_2) \sim \mathcal{M}\left(n_C; \frac{(1-f_0) r_0}{1-K_p}, \frac{(1-f_1) r_1}{1-K_p}, \frac{(1-f_2) r_2}{1-K_p}\right).$$

In such a context, H_0: $\{RR_2 = 1\}$ and H_1: $\{RR2 \neq 1\}$. Once the alternative hypothesis is explicit, power can be calculated using one of the following frameworks to approximate the distribution of $\mathcal{S}$ under H_1.

Monte-Carlo estimation

As long as it is possible to generate a case-control sample $X = \begin{Bmatrix} D_0 & D_1 & D_2 \\ C_0 & C_1 & C_2 \end{Bmatrix}$. under H_1, it is very easy to estimate the power. We first draw N samples denoting $x^{(i)}$ the i^{th} sample. From this sample we get N statistics $s^{(1)}, \ldots, s^{(N)}$ from which we get the estimation of the power:

$$\widehat{\pi}(\alpha) = \frac{\sharp\left\{s^{(i)} \geq t_\alpha\right\}}{N}.$$

This well-known method is often very easy to perform and is consequently widely used, particularly in the field of statistical genetics when alternative distributions are hard to calculate analytically (Longmate, 2001). But such an approach may generally require a lot of time to reach a given level of precision. As $\widehat{\pi}$ is distributed according to a binomial distribution, using the central limit theorem

we find that $\widehat{\pi} \sim \mathcal{N}(\pi, \pi(1-\pi)/N)$, which gives the following 95% confidence interval:

$$\left[\widehat{\pi} - 1.96\frac{\sqrt{\widehat{\pi}(1-\widehat{\pi})}}{\sqrt{N}};\quad \widehat{\pi} + 1.96\frac{\sqrt{\widehat{\pi}(1-\widehat{\pi})}}{\sqrt{N}}\right].$$

Consequently the precision of the power estimate increases with speed $1/\sqrt{N}$. One could remark that the same method can be used to estimate the threshold of tests involving statistics for which the distribution under H_0 is not easily available (*e.g.* meta-statistics).

Asymptotic non-centrality parameter

Mitra (1958) demonstrated that, under H_1, the asymptotic distribution of a chi-square frequency test applied to a $2 \times c$ contingency table follows a non-central chi-square distribution $\chi'^2(k, \lambda)$, where k is the degree of freedom and λ the non-centrality parameter, such that

$$\lambda = N_1 N_2 \times \sum_{j=1}^{c} \frac{(p_{1j} - p_{2j})^2}{N_1 p_{1j} + N_2 p_{2j}},$$

with p_{ij} the frequency of case ij and N_1, N_2 the total counts of the first and second row. Mitra derived the asymptotic power for the test:

$$\pi(\alpha) \underset{\infty}{\rightarrow} 1 - \chi'^2_{1-t_\alpha}(k, \lambda).$$

Given the expression of the non-centrality parameter, this approach can be adapted to any statistic following a chi-square distribution under H_0 (see below for the particular case of the trend test) and is appropriate when sample sizes are large enough. It has recently been presented as an appealing and fast way to approximate power in association studies (Sham et al. 2000; Gordon et al. 2002; Kang *et al.* 2004).

Delta-Method

The Delta-Method is used to approximate the distribution of $\mathcal{S}$ with X. The multinomial distribution of X (derived from the genetic model) is asymptotically distributed according to a Gaussian distribution $\mathcal{N}(M, \Sigma)$. Using an order-1 Taylor development of $\mathcal{S} = f(X)$ around M we hence approximate $\mathcal{S}$ by:

$$\mathcal{S} \simeq f(M) + {}^t(X - M) \times \nabla f(M),$$

where ^{t} is the transpose operator and ∇f is the gradient of f. This 1-order development allows us to approximate the distribution of $\mathcal{S}$ by a Gaussian distribution $\mathcal{N}(m, \sigma^2)$, with $m = f(M)$ and $\sigma^2 = {}^t\nabla f(M) \times \Sigma \times \nabla f(M)$. Then we have:

$$\pi(\alpha) \underset{\infty}{\rightarrow} 1 - \Phi\left(\frac{t_\alpha - m}{\sigma}\right)$$

where Φ is the cumulative distribution function (CDF) of a Gaussian variable with zero mean and a variance of one. Of course, the closer the distribution under H_1 is to a Gaussian distribution, the better will be this 1-order approximation.

For cases where the Gaussian distribution of the statistic under H_1 is not realistic, we propose to use a order-2 Taylor expansion around M. We hence get a more precise approximation based on the distribution of a quadratic form in normal variables (QFNV):

$$\mathcal{S} \simeq f(M) + {}^t(X - M) \times \nabla f(M) + \frac{1}{2}{}^t(X - M) \times \nabla^2 f(M) \times (X - M),$$

where $\nabla^2 f$ is the Hessian of f.

In the case of the first order development, the computation of power only requires evaluating the CDF of a normal distribution. With the second order development, however, the distribution of $\mathcal{S}$ is approximated by a combination of chi-squares and the CDF is not straighforward to derive. Technical details can be found in Appendix 1 and derivations of the distribution for the statistics considered are available at: http://stat.genopole.cnrs.fr/software/dm2.

Application

Statistics considered

Here we consider four statistics.

(i) The *genotypic test* compares genotypic frequencies between affected and unaffected subjects by using the Pearson's chi-square statistic:

$$\mathcal{S}_G = \sum_{i=0}^{2} \frac{\left(D_i - \frac{n_D \times n_i}{n}\right)^2}{\frac{n_D \times n_i}{n}} + \frac{\left(C_i - \frac{n_C \times n_i}{n}\right)^2}{\frac{n_C \times n_i}{n}} \underset{H_0}{\sim} \chi^2(2),$$

From the formula given in (2.2.3) and the parameters of the genetic model (2.2.1), we can derive

the non-centrality parameter for this statistic:

$$\lambda_G = n_D n_C \times \sum_{i=0}^{2} \frac{\left(\frac{f_i r_i}{Kp} - \frac{(1-f_i)r_i}{1-Kp}\right)^2}{n_D \frac{f_i r_i}{Kp} + n_C \frac{(1-f_i)r_i}{1-Kp}}$$

and $\mathcal{S}_G \widetilde{H_1} \chi^2(2, \lambda_G)$

(ii) Another test based on genotypes is the *Cochran-Armitage test for trends* (Armitage, 1995). It measures a linear trend in proportions weighted by a dose effect score x_i associated to each column with x_i corresponding to the number of susceptibility allele:

$$\mathcal{S}_T = \frac{n.[n.(D_1 + 2D_2) - n_D.(n_1 + 2n_2)]^2}{n_D n_C.[n.(n_1 + 4n_2) - (n_1 + 2n_2)^2]} \widetilde{H_0} \chi^2(1).$$

For this particular case (trend test with three categories) Gordon *et al.* (2005) derived the expression of the non-centrality parameter, based on previous work of Chapman & Nam (1968). With our notation, it comes down to:

$$\lambda_T = n_D n_C \times \frac{\left[\sum x_i \left(\frac{(1-f_i)r_i}{1-Kp} - \frac{f_i r_i}{Kp}\right)\right]^2}{\sum x_i^2 \left(n_D \frac{f_i r_i}{Kp} + n_C \frac{(1-f_i)ri}{Kp}\right) - \frac{\left[\sum x_i \left(n_D \frac{f_i r_i}{Kp} + n_C \frac{(1-f_i)ri}{Kp}\right)\right]^2}{n}}$$

and $\mathcal{S}_T \widetilde{H_1} \chi^2(1, \lambda_T)$

Note that chi-square approximations are appropriate when sample sizes are in accordance with Cochran's condition (each expected cell count > 5 -, Cochran, 1952).

(iii) Another strategy is to combine simple statistics via meta-statistics with the hope of gaining power. Nevertheless, the actual null hypothesis tested is not explicit and distributions of such statistics (under H_0 or H_1) are not easy to assess out of Monte-Carlo simulations. Our aim considering $\mathcal{S}_\Sigma = \mathcal{S}_G + \mathcal{S}_T$ and $\mathcal{S}_\Pi = \mathcal{S}_G \times \mathcal{S}_T$ is to assess the efficiency of competing approaches to handle their power computation.

Simulations

Simulations are performed using the susceptibility allele frequency (p) as a factor of variation. All simulations are considered for a prevalence $K_p = 0.05$, $n_D = n_C = 500$ and the four MOIs ($RR_2 = 1.5$). Each Monte-Carlo estimate of power is carried out on the basis of $N = 10,000$ simulations, and is considered as a reference to compare with the other approaches. Using this approach to compute a power $\widehat{\pi}$, we get a 95% confidence interval of radius $0.0196\sqrt{\widehat{\pi}(1-\widehat{\pi})}$ centered on $\widehat{\pi}$. For example, this radius gives 0.588% for $\widehat{\pi} = 10\%$ (or 90%), 0.784% for $\widehat{\pi} = 20\%$ (or 80%) and is always smaller than 0.98% (case $\widehat{\pi} = 50\%$).

Results

Results concerning $\mathcal{S}_G$ and $\mathcal{S}_T$ are compiled in Figure 1. For the set of parameters considered, the additive and multiplicative models give very close results so we have displayed them only for the additive, recessive and dominant models.

The non-central chi-square approach (NC) is fully adapted to chi-square distributed statistics and hence gives accurate estimates of power. As non-central chi-square distributions are particular cases of QFNV, the order-2 Delta-Method (DM2) unsurprisingly also gives good results. By comparison, the order-1 Delta-Method (DM1) underestimates the power in the two cases. This underscores the fact that the Gaussian approximation under H_1 made by this approach is not realistic. However it provides better estimates for the trend test than for the genotypic test. This variation is due to the fact that the distribution of $\mathcal{S}_T$ under H_0 and H_1 is closer to a Gaussian distribution - that requires this approach - than $\mathcal{S}_G$. For instance, the expected value for the Wilk-Shapiro statistic test for normality is 0.69 for one degree-of-freedom chi-square distributed samples and 0.81 for two degrees-of-freedom chi-square distributed ones.

In the literature it has been suggested that factors such as the ratio of cases to controls, minor allele frequency and total sample size affect the accuracy of the analytic power calculations (Ji *et al.* 2005). We investigated such effects considering the genotypic and trend tests for the four MOIs and values of (0.04, 0.2, 1), (0.2,0.3,0.4,0.5)

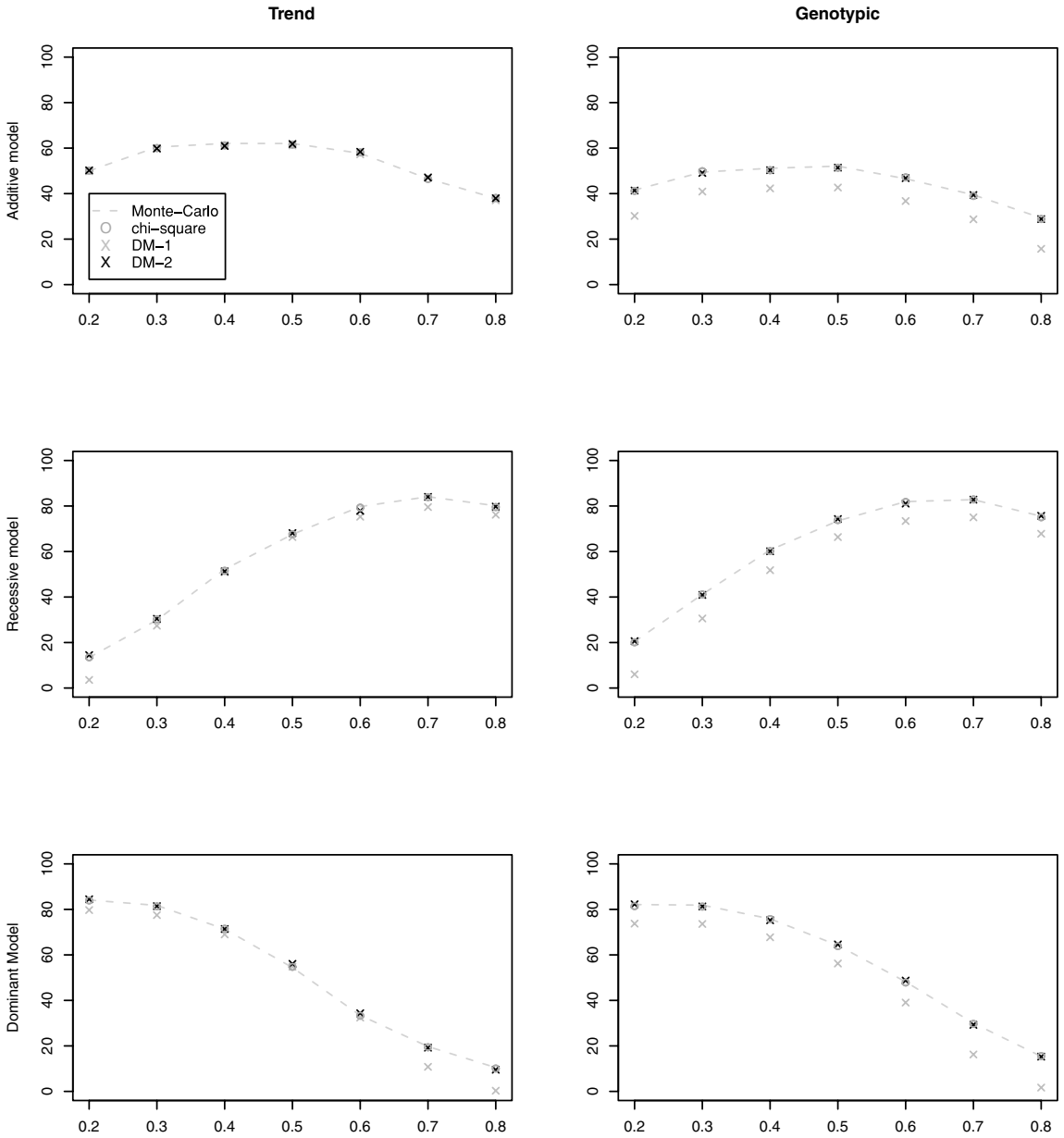


Figure 1 Power estimation (at the 5% significance level) for the trend and genotypic tests according to the allele frequency (p).

and (40,200,1000) for the ratio, the minor allele frequency and the total sample size, respectively. Normalized absolute differences have been computed for the NC and DM2 calculations (data not shown). However, we did not observe any clear effect of these factors on the accuracy of the analytic calculations.

Figure 2 presents the results for the two meta-statistics ($\mathcal{S}_\Sigma$ and $\mathcal{S}_\Pi$). In these cases, the non-central chi-square approach is not applicable. Even if $\mathcal{S}_\Sigma$ is the sum of two chi-square distributed statistics, $\mathcal{S}_G$ and $\mathcal{S}_T$ are not independent and hence $\mathcal{S}_\Sigma$ is not merely distributed according to a three degrees-of-freedom chi-square distribution under H_0. DM1 still badly estimates power. DM2 is really efficient to treat $\mathcal{S}_\Sigma$. As previously underlined, a linear combination of (dependent or not) chi-square distributed statistics is a QFNV which explains that DM2 works well on $\mathcal{S}_\Sigma$. Nevertheless DM2 does not manage to assess the

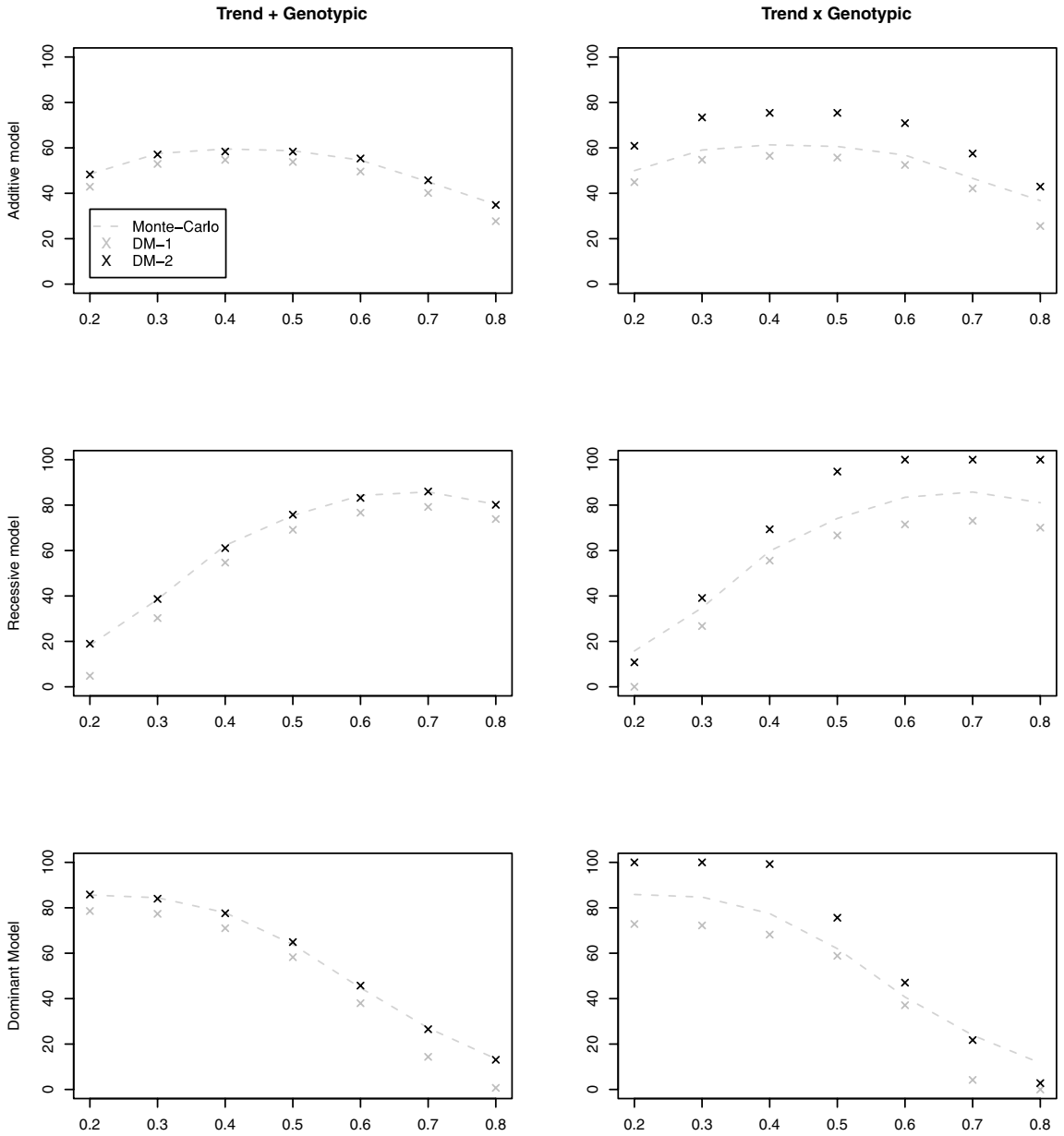


Figure 2 Power estimation (at the 5% significance level) for the meta-statistics according to the allele frequency (p).

power of $\mathcal{S}_\Pi$. We can imagine that such a product of chi-square distributed statistics would have required the use of the Delta-Method to a higher order, for which the determination of the CDF would have been numerically very expensive and unrealistic in practice.

In terms of comparison between the four strategies considered, $\mathcal{S}_T$, $\mathcal{S}_G$, $\mathcal{S}_\Sigma$ and $\mathcal{S}_\Pi$ (Figure 3), meta-statistics power estimates mainly lie between trend and genotypic ones and hence do not clearly represent a better alternative to single-statistics. However, they do more than merely averaging power estimates of single statistics, and hence can appear as a clever alternative to combine efficiency according to the model.

Discussion

Studying power is an important tool in statistics to compare the efficiency of different tests or to help design a study. With the accumulation of new analysis methods that have recently arisen from the accumulation of

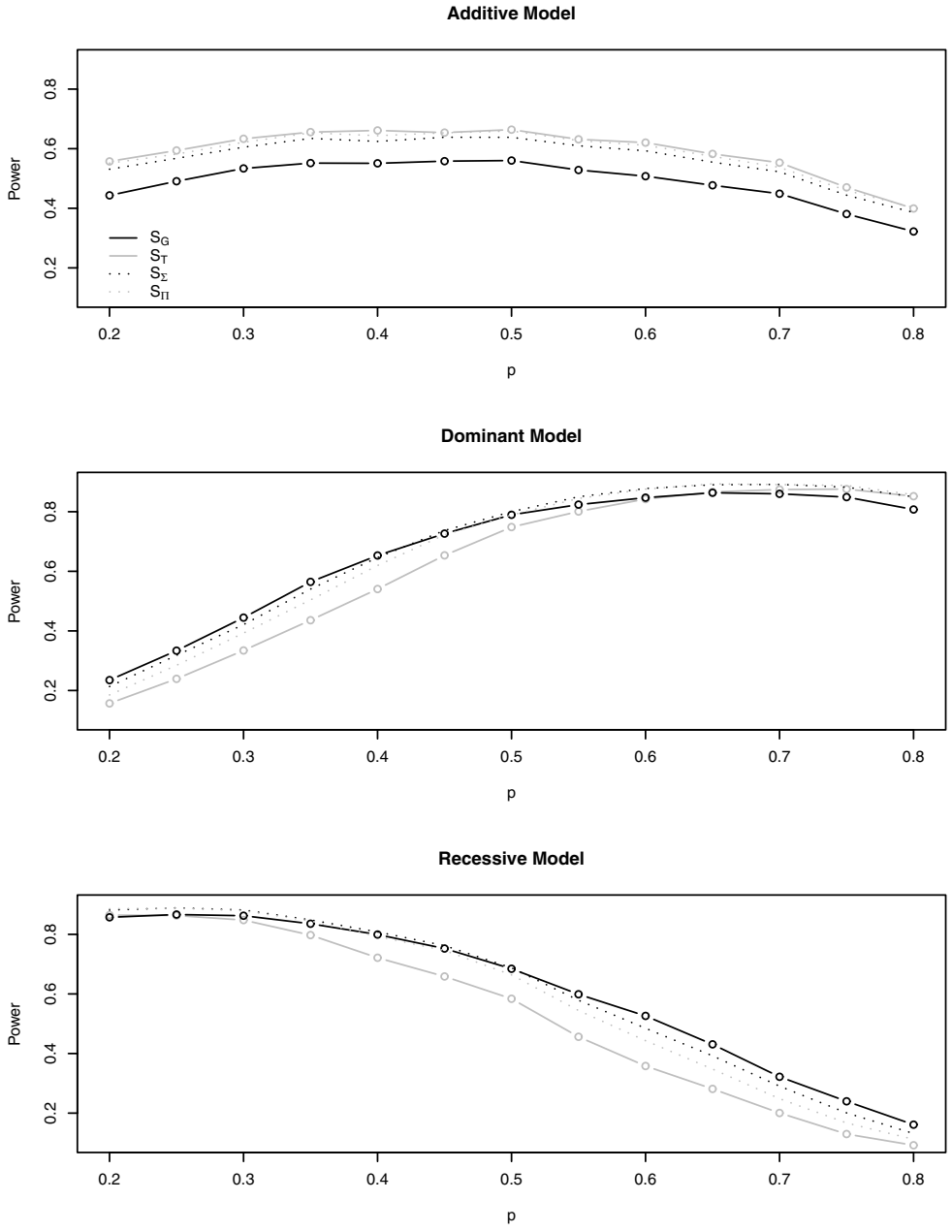


Figure 3 Power comparison: this figure compares the power of the four statistics $\mathcal{S}_G$, $\mathcal{S}_T$, $\mathcal{S}_\Sigma$ and $\mathcal{S}_\Pi$ according to the allele frequency (p). It is done for the additive, recessive and dominant models. Power is computed by Monte-Carlo at the 5% significance level.

large-scale data, statistical genetics does not escape this rule. In this article we focus on the computation of power in the context of simple-marker analyzes *via* the genotypic and trend statistics, as well as simple combinations of them.

Easy to implement, Monte-Carlo simulations are often the preferred approach to compute power estimations. Nevertheless they are computationally expensive since the precision of the estimates is directly dependent on the number of simulations performed. In particular the length of the confidence interval decreases with $1/\sqrt{N}$, and hence evolves quite slowly with N. Computing power through the non-centrality parameter is logically well adapted for statistics distributed

according to a chi-square distribution under H_0. The order-1 Delta-Method is based on a Gaussian distribution of the statistic. As a result it is not efficient in the situation considered here. In the literature approaches based on the order-1 Delta-Method have been successfully developed (Slager & Schaid, 2001; Jackson et al. 2002) to compute accurate power approximations for allelic and trend tests (Slager & Schaid, 2001; Jackson et al. 2002). The required Gaussian distributions of the statistics were obtained by the authors considering $\mathcal{Z} \sim \mathcal{N}(0, 1)$ such that $(\mathcal{Z})^2 = \mathcal{S}$ instead of $\mathcal{S}$ directly (as we have done here). This approach provides very good power approximations. However, its application is restrained to z-scores or by extension to 1 degree-of-freedom chi-square distributed statistics. It is hence less general than the non-central chi-square approach. To go further, we introduce the use of the order-2 Delta-Method. This approach provides good estimates and can be used to treat simple statistics and linear combinations of them, which is an advantage over other approaches. Besides a less straightforward CDF evaluation, it represents much less computionally expensive alternative to Monte-Carlo simulations, more general than the non-central chi-square framework and more accurate than the order-1 Delta-Method.

This work has been restricted to the study of the trend and genotypic tests under alternatives that differ in the susceptibility allele frequency and the MOI only. However, our conclusions can easily be extended to other simple-marker tests (Hardy-Weinberg and allelic tests, for instance) based on more complicated meta-statistics, and applied to more elaborate alternative models taking, for instance, the coefficient of consanguinity, linkage disequilibrium and genotyping errors into account.

Even if they fail to provide greater power estimates than single statistics, meta-statistics do not suffer from substantial power loss when compared with the best single statistic for each of the situations considered in this work. We thus suggest that meta-statistics may provide a useful means for combining such tests.

If they fail to provide better results than single-statistics, meta-statistics do not present a sensible loss of power compared to the best simple statistic in each situation considered, and hence appear to be a clever possible way to combine such tests.

Acknowledgements

The authors would like to thank Bernard Prum's team as well as members of the Serono Pharmaceutical Institute, in particular Jérôme Wojcik and Hiroaki Tanaka for encouraging this work.

References

Armitage, P. (1995). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

Chapman, D. and Nam, J. (1968). Asymptotic power of chi-square tests for linear trends in proportions. *Biometrics* **24**, 315–327.

Cochran, W. (1952). The chi-square test of goodness of fit. *Annals of Human Genetics* **23**, 315–345.

Davies, R. B. (1973). Numerical inversion of characteristic function. *Biometrika* **60**, 481–482.

Davies, R. B. (1980). The distribution of a linear combination of chi-square random variables. *Appl. Stat* **29**, 323–333.

Gordon, D., Finch, S. J., Nothnagel, M. and Ott, J. (2002). Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotid polymorphisms. *Human Heredity* **54**, 22–33.

Gordon, D., Haynes, C., Blumenfeld, J. and Finch, S. (2005). Pawe-3d: visualizing power for association with error in case-control genetic studies of complex traits. *Bioinformatics* **21**, 3935–3937.

Jackson, M., Genin, E., Knapp, M. and Ji. E. (2002). Accurate power approximation for X-tests in case-control association studies of complex disease genes. *Annals of Human Genetics* **66**, 307–321.

Ji, F., Yang, Y., Haynes, C., Finch, S. and Gordon, D. (2005). Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors. *Statistics Applied to Genetic and Molecular Biology* 4.

Kang, S. J., Gordon, D. and Finch, S. J. (2004). What snp genotyping errors are most costly for genetic association studies. *Genetic epidemiology* **26**, 132–141.

Longmate, J. A. (2001). Complexity and power in case-control association studies. *American Journal Human Genetics* **68**, 1229–1237.

Lu, Z.-H. and King, M. L. (2002). Improving the numerical techniqie for computing the accumulated distribution of a quadratic form in normal variables. *Econometric review* **21**, 149–165.

Mitra, S. (1958). On the limiting power function of the frequency chi-square test. *Annals of Mathematical Statistics* **29**, 1221–1233.

Risch, N. (2000). Searching for genes in complex diseases: lessons from systemic lupus erythematosus. *American society for clinical investigation* **105**, 1503–1506.

Sham, P., Cherny, S., Purcell, S. and Hewitt, J. (2000). Power of linkage versus association analysis of quatitative traits, by use of variance components models, for sibship data. *American Journal Human Genetics* **66**, 1616–1630.

Slager, S. and Schaid, D. (2001). Case-control studies of genetic markers: power and sample size approximations for armitage's test for trend. *Human Heredity* **52**, 149–153.

Zhao, H. (2000). Family-based association studies. *Statistical Methods in Medical Research* **9**, 563–587.

Appendix A: Quadratic Form in Normal Variables

In this appendix we propose to recall the definition of this distribution and explain how it is possible to compute its cumulative distribution function (CDF).

Definition 1 If $X \sim \mathcal{N}(\mu, \Sigma)$ *is a dimension* $d \geq 1$ *(column) vector of normal variables we call*

$$Q = A + BX + X'CX$$

a quadratic form in normal variables *(QFNV) of dimension d* with *parameters* $A \in \mathbb{R}$, $B \in \mathbb{R}^{(1,d)}$ *and* $C \in \mathbb{R}^{(d,d)}$ and with $\mu \in \mathbb{R}^{(d,1)}$ *and* $\Sigma \in \mathbb{R}^{(d,d)}$ *are mean and covariance matrix of the normal variables.*

In particular, a linear combination of (central or not, independent or not) chi-square is a QFNV.

For non-degenerate parameters, it is possible to express a QFNV as a linear combination of independent non-central chi-square distribution (Lu and King, 2002). Namely:

Proposition 2 *For any QFNV Q with non-singular covariance matrix* Σ *and matrix* C, $1 \leq n \leq d$, $\lambda_j \in \mathbb{R}$, $d_j \in \mathbb{N}^*$ *and* $\nu_j > 0$ *(for all* $1 \leq j \leq n$*) such as*

$$Q = K + \sum_{j=1}^{n} \lambda_j \chi_j$$

where $K \in \mathbb{R}$ *and* $\chi_j \sim \chi^2(d_j, \nu_j)$ *are independent chi-square variables with* d_j *degrees of freedom and* ν_j *non centrality* parameters.

Proof. We first factorize Q in

$$Q = \underbrace{\left(A - \frac{BC^{-1}B'}{4}\right)}_{K} + \underbrace{\left(X + \frac{C^{-1}B'}{2}\right)' C \left(X + \frac{C^{-1}B'}{2}\right)}_{Y'CY}$$

where $Y \sim \mathcal{N}(\tilde{\mu}, \Sigma)$ with $\tilde{\mu} = \mu + C^{-1}B'/2$. We consider then the linear transformation $Z = \Sigma^{-1/2}\, Y$ so $Z \sim \mathcal{N}(\Sigma^{-1/2}\tilde{\mu}, I)$ and

$$Q = K + Z' \underbrace{\Sigma^{-1/2} C \left(\Sigma^{1/2}\right)'}_{\tilde{C}} Z$$

We then consider the orthogonal matrix P of the eigenvector of $\tilde{C}$ and denote by $D = P'\tilde{C}P$ the diagonal matrix of the corresponding eigenvalues. With $W = P^{-1}\, Y$ we get

$$Q = K + W' DW$$

with $W \sim \mathcal{N}(P^{-1}\Sigma^{-1/2}\tilde{\mu}, I)$. For $1 \leq j \leq n$, we denote by λ^j the (distinct) eigenvalue and by d_j their orders of multiplicity (one should note that they are also those of $C\Sigma$ or ΣC). Finally, we consider $\nu_j = \sum_{q=1}^{d_j} \gamma_{j,q}^2$ where $\gamma_{j,q}$ are the elements of $P^{-1}\Sigma^{-1/2}\tilde{\mu}$ corresponding to the same eigenvalue λ_j and the result is established.

From now, we hence focus of the numerical CDF evaluation of

$$Q = \sum_{j=1}^{n} \lambda_j \chi^2(d_j, \nu_j)$$

a linear combination of independent and non central chi-square distributions.

A numerical inversion of the characteristic function is then possible, resulting through truncation and trapezoidal integration (Davies, 1973; Davies, 1980) in the following formula

$$\mathbb{P}(Q < c) = \frac{1}{2} - \sum_{m=0}^{M} \left(\frac{\sin\{\theta_c[(m + 0.5)\Delta]\}}{\pi(m + 0.5)\gamma[(m + 0.5)\Delta]} \right)$$

where Δ is the (small) step interval, M the (large) number of step intervals, and $U = (M + 0.5)\delta$ the truncation value. The functions θ_c and γ are given by

$$\theta_c(u) = \sum_{j=1}^{n} \left[\frac{d_j}{2} \tan^{-1}(2u\lambda_j) + \nu_j u \lambda_j (1 + 4u^2\lambda_j^2)^{-1} \right] - c\,u$$

and

$$\gamma(u) = \prod_{j=1}^{n} (1 + 4u^2\lambda_j^2)^{d_j/4} \exp\left(2u^2 \sum_{j=1}^{n} \frac{\nu_j \lambda_j^2}{1 + 4u^2\lambda_j^2} \right)$$

The numerical evaluation of the CDF using this formula leads to an error of truncation ε_T depending on the truncation bound U, and to an error of integration ε_I depending on the step interval Δ. There exist many concurrent ways to choose both these values and Lu & King (2002) provide a complete review of them.

Annexes

Forme Quadratique

Nous nous proposons ici de rappeler la définition de la distribution d'une forme quadratique en variables normales et d'expliquer comment il est possible de calculer sa fonction de répartition (CDF).

Définition

Soir $X \sim \mathcal{N}(\mu, \Sigma)$ un vecteur colonne de variables normales de dimension $d \geqslant 1$ avec $\mu \in \mathbb{R}^{(d,1)}$ et $\Sigma \in \mathbb{R}^{(d,d)}$ le vecteur moyenne et la matrice de covariance des variables normales respectivement. On définit alors par :

$$Q = A + BX + X'CX,$$

une forme quadratique en variables normales (QFNV) de dimension d et de paramètres $A \in \mathbb{R}$, $B \in \mathbb{R}^{(1,d)}$ et $C \in \mathbb{R}^{(d,d)}$. En particulier, une combinaison linéaire de χ^2 (décentrés ou non, indépendants ou non) est une QFNV.

Pour peu que les paramètres A, B et C ne soit pas dégénérés, il est possible d'exprimer une QFNV comme une combinaison de χ^2 décentrés indépendants (Lu and King, 2002). A savoir :

Proposition : Pour toute QFNV Q avec une matrice de covariance Σ non-singulière et toute matrice C, il existe $1 \leqslant n \leqslant d$, $\lambda_j \in \mathbb{R}$, $d_j \in \mathbb{N}^*$ et $\nu_j > 0$ (pour tout $1 \leqslant j \leqslant n$) tels que :

$$Q = K + \sum_{j=1}^{n} \lambda_j \chi_j,$$

où $K \in \mathbb{R}$ et $\chi_j \sim \chi^2(d_j, \nu_j)$.

Preuve : Dans un premier temps on factorise Q dans :

$$Q = \underbrace{\left(A - \frac{BC^{-1}B'}{4}\right)}_{K} + \underbrace{\left(X + \frac{C^{-1}B'}{2}\right)' C \left(X + \frac{C^{-1}B'}{2}\right)}_{Y'CY},$$

où $Y \sim \mathcal{N}(\tilde{\mu}, \Sigma)$ avec $\tilde{\mu} = \mu + C^{-1}B'/2$. On considère ensuite la transformation linéaire $Z = \Sigma^{-1/2}Y$ de sorte que $Z \sim \mathcal{N}(\Sigma^{-1/2}\tilde{\mu}, I)$ et

$$Q = K + Z' \underbrace{\Sigma^{-1/2} C \left(\Sigma^{1/2}\right)'}_{\tilde{C}} Z.$$

Soit la matrice orthogonale P des vecteurs propres de $\tilde{C}$ et $D = P'\tilde{C}P$ la matrice diagonale correspondant aux valeurs propres. Avec $W = P^{-1}Y$ nous obtenons :

$$Q = K + W'DW,$$

avec $W \sim \mathcal{N}(P^{-1}\Sigma^{-1/2}\tilde{\mu}, I)$. Pour $1 \leqslant j \leqslant n$, l'on dénote par λ^j les valeurs propres distinctes et par d_j leur multiplicité. Pour finir, l'on considère $\nu_j = \sum_{q=1}^{d_j} \gamma_{j,q}^2$ où $\gamma_{j,q}$ sont les éléments de $P^{-1}\Sigma^{-1/2}\tilde{\mu}$ correspondant aux mêmes valeurs propres λ^j et le résultat est établi.

Estimation de la CDF

L'inversion numérique de la fonction caratéristique de la QFNV est possible en appliquant une troncature et une intégration trapézoïdale (Davies 1973, Davies 1980) à la formule suivante :

$$\mathbb{P}(Q < c) = \frac{1}{2} - \sum_{m=0}^{M} \left(\frac{\sin\{\theta_c[(m + 0.5)\Delta]\}}{\pi(m + 0.5)\gamma[(m + 0.5)\Delta]} \right)$$

où Δ est le (petit) intervalle, M le (grand) nombre d'intervalles, et $U = (M + 0.5)\delta$ la valeur de troncature. Les fonctions θ_c et γ sont données par :

$$\theta_c(u) = \sum_{j=1}^{n} \left[\frac{d_j}{2} \tan^{-1}(2u\lambda_j) + \nu_j u\lambda_j (1 + 4u^2\lambda_j^2)^{-1} \right] - cu$$

et

$$\gamma(u) = \prod_{j=1}^{n} (1 + 4u^2\lambda_j^2)^{d_j/4} \exp\left(2u^2 \sum_{j=1}^{n} \frac{\nu_j \lambda_j^2}{1 + 4u^2\lambda_j^2} \right).$$

L'évaluation numérique de la CDF de cette façon aboutit à une erreur de troncature ε_T (liée à la limite de troncature U) ainsi qu'à une erreur d'intégration ε_I (lié à l'intervalle Δ). Il existe un certains nombre de stratégies pour choisir ces valeurs et Lu and King (2002) fournit une *review* complète.

X_A et X_T sont asymptotiquement équivalents à l'équilibre d'Hardy-Weinberg

Lorsque l'on fait le rapport entre les deux statistiques, l'on obtient :

$$\frac{X_A}{X_T} = \frac{4N_0N_2 - N1^2}{(N_1 + 2N_2)(N_1 + 2N_0)} + 1 = \frac{A}{B} + 1,$$

avec N_0, N_1 et N_2 des variables aléatoires de distribution multinomiale :

$$(N_0, N_1, N_2) \sim \mathcal{M}(N, p_0, p_1, p_2),$$

et qui peuvent être approchées par une distribution multinormale de paramètres connus. Par ailleurs, p_0, p_1 et p_2 sont ici les proportions génotypiques dans la population combinée cas-témoin que l'on suppose également à l'équilibre d'Hardy-Weinberg avec $p_0 = p^2$, $p_1 = 2p(1-p)$ et $p_2 = (1-p)^2$.

Pour montrer que X_A et X_T sont asymptotiquement équivalents dans ces conditions, il faut au moins montrer que le quotient de A avec B tend vers 0 lorsque n tend vers l'infini :

$$\lim_{N \to +\infty} \frac{A}{B} = 0.$$

En posant $N_0 = Np_0 + Y_0$, $N_2 = Np_2 + Y_2$ et $N_1 = Np_1 - Y_0 - Y_2$ où Y_0 et Y_2 sont des variables normales centrées, on a :

$$\begin{aligned} A &= 4N(Y_0(1-p) + Y_2p) - (Y_0 - Y_2)^2, \\ B &= 4N^2p^2 - (Y_0 - Y_2)^2. \end{aligned}$$

Soient $D_N = Y_0 - Y_2$ et $S_N = Y_0(1-p) + Y_2p$. On peut montrer que $D_N = \sqrt{N}D$ et $S_N = \sqrt{N}S$ avec :

$$\begin{aligned} D &\sim \mathcal{N}(0, p_1), \\ S &\sim \mathcal{N}(0, p_0p_2). \end{aligned}$$

On pose :

$$Q = \frac{\frac{A}{N\sqrt{N}}}{\frac{B}{N^2}}.$$

Avec le théorème de Slutzky (Billingsley 1968), on peut montrer que Q converge en loi vers une normale $\mathcal{N}(0, \frac{p_2}{p_0})$. On peut en déduire que :

$$\lim_{N\to+\infty} \frac{A}{B} = \lim_{N\to+\infty} \frac{1}{\sqrt{N}} \times Q = 0,$$

ce qu'il fallait démontrer.

Justification des conditions d'application du test d'Hardy-Weinberg

Condition 1 : la population générale doit être à l'équilibre

Un déséquilibre observé chez les cas peut-être dû à trois raisons principales : **(i)** la population générale n'est pas à l'équilibre, **(ii)** le déséquilibre observé est obtenu par chance, **(iii)** le déséquilibre est lié à une association entre le marqueur considéré et la maladie. En contrôlant le fait que le déséquilibre soit obtenu par chance, pour faire l'hypothèse qu'il est dû à la maladie, l'on doit dans un premier temps exclure celle qu'il soit dû à un déséquilibre dans la population générale.

Condition 2 : la pénétrance est incomplète

$$\begin{aligned}
\mathcal{F}_D = 0 &\Leftrightarrow p_{D_2} - (p_{D_A})^2 = 0 \\
&\Leftrightarrow \frac{p_A^2 f_2}{K_p} - \left[p\frac{f_2 p_A + f_1(1-p_A)}{K_p}\right]^2 = 0 \\
&\Leftrightarrow \frac{p_A^2 K_p f_2 - p_A^2[f_2 p_A + f_1(1-p_A)]^2}{K_p^2} = 0 \\
&\Leftrightarrow \frac{(1-p_A)^2 p_A^2 (f_0 f_2 - f_1^2)}{K_p^2} = 0 \\
&\Leftrightarrow f_0 f_2 - f_1^2 = 0 \\
&\Leftrightarrow f_0 = f_1 = f_2 = 1 \text{ OR } f_0 = f_1 = 0
\end{aligned}$$

Condition 3 : le mode de transmission ne doit pas être multiplicatif

$$\begin{aligned}\mathcal{F}_D = 0 &\Leftrightarrow p_{D_2} - (p_{D_A})^2 = 0 \\ &\Leftrightarrow f_0 f_2 - f_1^2 = 0 \\ &\Leftrightarrow f_0 f_0 \times RR_2 - (f_0 \times RR_1)^2 = 0 \\ &\Leftrightarrow f_0^2 RR_2 - f_0^2 RR_1^2 = 0 \\ &\Leftrightarrow RR_2 = RR_1^2\end{aligned}$$

Sous un mode de transmission multiplicatif, les génotypes sont sélectionnés proportionnellement au produit des proportions alléliques. Par conséquent, on ne s'attend pas à observer chez les cas un déséquilibre d'Hardy-Weinberg généré par la maladie et plus le mode de transmission s'éloigne du modèle multiplicatif, et plus la quantité de déséquilibre susceptible d'être générée par la maladie est importante.

Introduction à l'algorithme EM

L'algorithme EM (*Expectation-Maximization*) est aujourd'hui un outil statistique très populaire pour traiter le problème de données incomplètes ou d'estimation de mélanges (Dempster et al 1977, McLachlan et Peel 2000). Il s'agit d'une procédure itérative de calcul de l'estimateur au maximum de vraisemblance en présence de données incomplètes ; c'est à dire que nous cherchons les paramètres du modèle considéré pour lesquels les données ont le plus de chance d'être observées. Chaque itération consiste en deux étapes : l'étape E et l'étape M.

Dans l'étape E, les données manquantes sont estimées sachant les données observées et une estimation courante des paramètres du modèle par l'espérance conditionnelle. Dans l'étape M, la fonction de vraisemblance est maximisée sous l'hypothèse que les données manquantes sont connues. Cet annexe est librement inspiré du tutorial proposé par Sean Borman[7].

Déduction

Soit $\mathbf{X}$ un vecteur aléatoire représentant les données observées. L'on cherche à déterminer θ tel que $\mathbb{P}(\mathbf{X}|\theta)$ est maximum. On parle alors d'estimation de θ au maximum de vraisemblance[8] avec $\mathbb{P}(\mathbf{X}|\theta)$ la fonction de vraisemblance de paramètre θ appliquée

[7] `http://www.seanborman.com/publications`

[8] note ML pour *Maximum Likelihood*

sur les observations $\mathbf{X}$. Afin d'estimer θ, on introduit généralement la fonction de log-vraisemblance définie par :

$$L(\theta) = \log \mathbb{P}(\mathbf{X}|\theta).$$

Comme $\log(x)$ est une fonction strictement croissante, la valeur de θ qui maximise $\mathbb{P}(\mathbf{X}|\theta)$ maximise également $L(\theta)$.

L'algorithme EM est une procédure itérative de maximisation de $L(\theta)$. Supposons qu'à la suite de la $n^{\text{ème}}$ itération, l'estimation de θ soit donnée par $\theta^{(h)}$; comme l'objectif est de maximiser $L(\theta)$, on cherche une estimation de θ telle que :

$$L(\theta) > L(\theta^{(h)}).$$

De façon équivalente, on cherche à maximiser la différence :

$$L(\theta) - L(\theta^{(h)}) = \log \mathbb{P}(\mathbf{X}|\theta) - \log \mathbb{P}(\mathbf{X}|\theta^{(h)}).$$

Pour le moment nous n'avons pas considéré de données non-observées. Soit $\mathbf{Z}$ un vecteur aléatoire représentant les données non-observées et $\mathbf{z}$ une réalisation de $\mathbf{Z}$. Il est alors possible d'exprimer $\mathbb{P}(\mathbf{X}|\theta)$ en fonction de $\mathbf{z}$:

$$\mathbb{P}(\mathbf{X}|\theta) = \sum_{\mathbf{z}} \mathbb{P}(\mathbf{X}|\mathbf{z}, \theta)\mathbb{P}(\mathbf{z}|\theta).$$

On peut alors réécrire $L(\theta) - L(\theta^{(h)})$ en fonction de $\mathbf{z}$:

$$L(\theta) > L(\theta^{(h)}) = \log \left(\sum_{\mathbf{z}} \mathbb{P}(\mathbf{X}|\mathbf{z}, \theta)\mathbb{P}(\mathbf{z}|\theta) \right) - \log \mathbb{P}(\mathbf{X}|\theta).$$

En utilisant l'inégalité de Jensen, on peut montrer que :

$$\log \sum \lambda_i x_i \geqslant \sum \lambda_i \log(x_i)$$

avec $\lambda_i \geqslant 0$ et $\sum \lambda_i = 1$. Ce résultat peut-être appliqué à l'équation précédente sous peine d'identifier les constantes λ_i. Ici elle peuvent être de la forme $\mathbb{P}(\mathbf{z}|\mathbf{X}, \theta^{(h)})$; comme il s'agit d'une probabilité, $\mathbb{P}(\mathbf{z}|\mathbf{X}, \theta^{(h)}) \geqslant 0$ et $\sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X}, \theta^{(h)}) = 1$ est requis. Nous avons donc en

introduisant ces constantes :

$$\begin{aligned}
L(\theta) - L(\theta^{(h)}) &= \log\left(\sum_{\mathbf{z}} \mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta)\right) - \log \mathbb{P}(\mathbf{X}|\theta) \\
&= \log\left(\sum_{\mathbf{z}} \mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta) \times \frac{\mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)})}{\mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)})}\right) - \log \mathbb{P}(\mathbf{X}|\theta) \\
&= \log\left(\sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \frac{\mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta)}{\mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)})}\right) - \log \mathbb{P}(\mathbf{X}|\theta) \\
&\geqslant \sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \log\left(\frac{\mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta)}{\mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)})}\right) - \log \mathbb{P}(\mathbf{X}|\theta) \\
&\geqslant \sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \log\left(\frac{\mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta)}{\mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)})\mathbb{P}(\mathbf{X}|\theta^{(h)})}\right) \\
L(\theta) &\geqslant L(\theta^{(h)}) + \sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \log\left(\frac{\mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta)}{\mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)})\mathbb{P}(\mathbf{X}|\theta^{(h)})}\right).
\end{aligned}$$

En posant :

$$l(\theta|\theta^{(h)}) \equiv L(\theta^{(h)}) + \sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \log\left(\frac{\mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta)}{\mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)})\mathbb{P}(\mathbf{X}|\theta^{(h)})}\right),$$

on obtient une fonction majorée par la fonction de vraisemblance $L(\theta)$. Par ailleurs on peut montrer que pour $\theta = \theta^{(h)}$, $l(\theta|\theta^{(h)})$ et $L(\theta)$ sont égales.

Notre objectif est de choisir une valeur de θ qui maximise $L(\theta)$. Nous avons montré que la fonction $l(\theta|\theta^{(h)})$ est majorée par $L(\theta)$ et que les deux sont égales pour une même valeur de θ. Par conséquent, tout θ qui accroît $l(\theta|\theta^{(h)})$ accroît également $L(\theta)$. De façon à réaliser le meilleur accroissement possible de $L(\theta)$, l'algorithme EM préconise de choisir θ tel que $l(\theta|\theta^{(h)})$ est maximum et noté $\theta^{(h+1)}$. Cette procédure est illustrée figure 4.1. Plus formellement nous avons :

$$\begin{aligned}
\theta_{n+1} &= \arg\max_{\theta}\{l(\theta|\theta^{(h)}\} \\
&= \arg\max_{\theta}\left\{L(\theta^{(h)}) + \sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \log \frac{\mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta)}{\mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)})\mathbb{P}(\mathbf{X}|\theta^{(h)})}\right\}
\end{aligned}$$

En retirant les termes constants par rapport à θ :

$$\begin{aligned}
\theta^{(h+1)} &= \arg\max_{\theta}\left\{\sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \log \mathbb{P}(\mathbf{X}|\mathbf{z},\theta)\mathbb{P}(\mathbf{z}|\theta)\right\} \\
&= \arg\max_{\theta}\left\{\sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \log \frac{\mathbb{P}(\mathbf{X},\mathbf{z},\theta)}{\mathbb{P}(\mathbf{z},\theta)} \frac{\mathbb{P}(\mathbf{z},\theta)}{\mathbb{P}(\theta)}\right\} \\
&= \arg\max_{\theta}\left\{\sum_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{X},\theta^{(h)}) \log \mathbb{P}(\mathbf{X},\mathbf{z}|\theta)\right\} \\
&= \arg\max_{\theta}\left\{\mathbb{E}_{\mathbf{z}|\mathbf{X},\theta^{(h)}}\{\log \mathbb{P}(\mathbf{X},\mathbf{z}|\theta)\}.\right\}
\end{aligned}$$

On voit apparaître ici les deux étapes de l'algorithme EM : **(i)** l'étape E qui détermine l'espérance conditionnelle $\mathbb{E}_{\mathbf{z}|\mathbf{X},\theta^{(h)}}\{\log \mathbb{P}(\mathbf{X}, \mathbf{z}|\theta)\}$ et **(ii)** l'étape M qui maximise l'expression par rapport à θ.

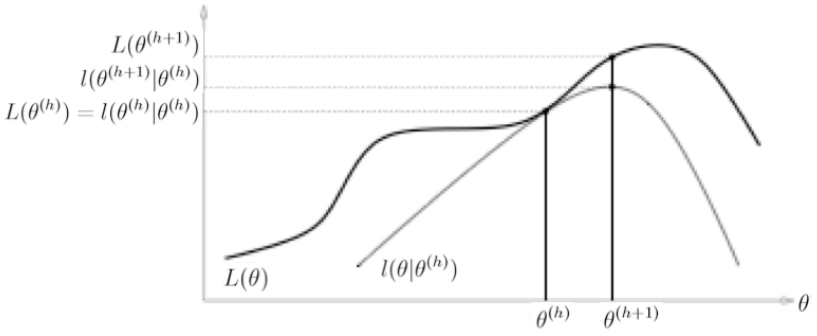


FIG. 4.1 – **Interprétation graphique d'une étape de l'algorithme EM :** la fonction $l(\theta|\theta^{(h)})$ est majorée pour la fonction de vraisemblance $L(\theta)$. Les fonctions sont égales pour $\theta = \theta^{(h)}$. L'algorithme EM choisit $\theta^{(h+1)}$ tel que $l(\theta|\theta^{(h)})$ est maximum. Comme $L(\theta) \geqslant l(\theta|\theta^{(h)})$, on est sûr en augmentant la valeur de $l(\theta|\theta^{(h)})$ d'augmenter par la même occasion $L(\theta)$ à chaque étape.

Convergence

Les propriétés de convergence de l'algorithme EM sont discutées en détails par McLachlan et Krishnan (1996). Ici nous discutons de la convergence générale de l'algorithme. La valeur $\theta^{(h+1)}$ est choisie de façon à accroître la fonction de vraisemblance $L(\theta)$. Quand l'algorithme atteint un point fixe pour un θ_n donné, cette valeur maximise alors $l(\theta)$. Comme L et l sont égales à $\theta^{(h)}$ (si L et l sont différentiables en ce point), alors $\theta^{(h)}$ est également un point de stationarité pour L. Néanmoins ce point n'est pas nécessairement un maximum global de la fonction de vraisemblance ; il est possible que l'algorithme converge vers un minimum local ou un point selle dans certains cas limites.

Résultats complémentaires sur le Score Local

Cette annexe est inspirée de la thèse présentée par David Robelin (2005).

Distribution asymptotique du Score Local

La principale idée de l'approche est de découper la séquence en blocs déterminés de manière aléatoire, et dont les distributions sont indépendantes et identiquement distri-

buées (iid). La manière dont ces blocs sont déterminés est présentée dans la suite. La méthode consiste à récrire le Score Local sur la séquence entière à l'aide d'une fonction définie sur chacun de ces blocs ; de cette façon, le Score Local s'exprime comme une fonction de variables aléatoires réelles et iid. La loi de ces variables aléatoires est étudiée sur le premier bloc.

Soit $\mathbb{S} = (\mathcal{S}_i)_{i=1,\ldots,n}$ une séquence de variables aléatoires iid et d'espérance négative. On note $\forall k \in [1,n]$, $M_k = \sum_{i=1}^{k} \mathcal{S}_i$ les sommes cumulées partielles. On note $\mathbf{M}$ le maximum des sommes cumulées partielles $\mathbf{M} = \sup_{k \geqslant 0} M_k$.

La démarche de Karlin et Dembo (1992) consiste à étudier la distribution de $\mathbf{M}$, pour ensuite en déduire la distribution du maximum des sommes cumulées partielles sur le premier bloc. Cette distribution permet ensuite de caractériser la distribution du plus grand Score Local. Les blocs sont définis à partir de temps de records négatifs de la marche aléatoire $\mathbb{M} = (M_k)_{k=1\ldots n}$. Les temps de record négatifs sont définis comme étant les temps d'arrêt suivants :

$$T_0 = 0 \text{ et } T_{k+1} = \inf\{i | i > T_k \text{ et } M_i \geqslant M_{T_k}\}.$$

Un temps de record correspond à un nouveau minimum dans la marche aléatoire $\mathbb{M}$. Comme l'espérance de $\mathcal{S}_i$ est négative, on a $\mathbb{P}(T_k < \infty) = 1$ et l'on note : $\mu = \mathbb{E}(T_1)$. Entre deux temps de records successifs, on s'intéresse aux variables aléatoires valant le maximum des sommes cumulées partielles sur un bloc :

$$Q_i = \max_{T_i \leqslant k \leqslant T_{i+1}} (M_k - M_{T_i}).$$

D'autre part, les trajectoires des sommes cumulées partielles entre deux temps de records successifs $\sum_k = (M_i, T_k \leqslant i \leqslant T_{k+1})_k$ sont iid. Ceci est dû à l'indépendance des variables aléatoires $\mathcal{S}_i$ et au fait qu'elles sont identiquement distribuées. Par conséquent, les maxima de ces trajectoires (Q_i) sont également des variables aléatoires iid. Le Score Local obtenu à un certain temps de record s'exprime comme le maximum de variables aléatoires iid :

$$H_{T_m} = \max_{0 \leqslant i \leqslant N_n} Q_i,$$

$\forall m \geqslant 0$ et où N_m est le nombre (aléatoire) de temps de records dans la séquence de taille n. Karlin et Dembo (1992) ont établi une approximation de la queue de la distribution du maximum des sommes cumulées partielles entre deux temps de records :

$$\lim_{y \to +\infty} \mathbb{P}(Q_1 > y) = Ce^{-\lambda y},$$

où λ est l'unique solution non nulle de l'équation $\mathbb{E}\left[e^{-x\mathcal{S}_1}\right] = 1$, et C est une constante dépendant de la loi des $\mathcal{S}_i$. Ils ont également établi que le domaine d'attraction de la loi de Q_i est la distribution de Gumbel ; autrement dit, la distribution du maximum des Q_i, c'est à dire H correctement normalisé, converge vers une distribution de Gumbel. On note :

$$\mathbf{M}_{N_n} = \max_{1 \leqslant i \leqslant N_n} Q_i.$$

D'après la théorie des valeurs extrêmes et le résultat de Karlin et Dembo (1992), il existe deux séquences α_k et β_k telles que :

$$\lim_{k\to+\infty} \mathbb{P}((\mathbf{M}_k - \alpha_k)/\beta_k < x) = e^{-e^{-x}},$$

avec :

$$\alpha_k = \frac{\log k + \log C}{\lambda}.$$
$$\beta_k = \frac{1}{\lambda},$$

La démonstration peut se trouver dans la thèse de David Robelin (2005). Cette proposition donne les constantes de normalisation en fonction du nombre de temps de records. Il reste à trouver ces constantes en fonction de la longueur de la séquence. On utilise pour cela la convergence presque sûre de N_n/n vers $1/\mu$. On peut donc affirmer que N_n se comporte presque sûrement comme n/μ. Cette observation conduit aux constantes de normalisation suivantes pour H :

$$\lim_{n\to+\infty} \mathbb{P}((H - \alpha_n)/\beta_n < x) = e^{-e^{-x}},$$

avec :

$$\alpha_n = \frac{\log n + \log(K)}{\lambda},$$
$$\beta_n = \frac{1}{\lambda},$$

avec $K = C/\mu$. La preuve est présentée dans la thèse de Sabine Mercier (1999). Nous venons donc de voir que ces constantes de normalisation conduisent à la convergence de la distribution du Score Local (H) d'une séquence vers une distribution de Gumbel. Ce résultat est lié à la convergence du maximum des Q_i vers la distribution de Gumbel.

Estimation pratique des constantes K et λ de la distribution de Gumbel

K et λ dépendent de la distribution de $\mathcal{S}_i$ et sont en pratique rarement faciles à déterminer. Une approche consiste à linéariser la fonction de répartition du Score Local obtenue par Monte-Carlo : $\hat{F}_H$. Cette technique repose sur la linéarisation de la distribution de Gumbel :

$$\lim_{n\to+\infty} \log(-\log(\mathbb{P}(H \geqslant x))) = \log K - \lambda x + \log n,$$

On en déduit directement que :

$$\log(-\log(\hat{F}_H(x))) = ax + b,$$

avec $a = -\lambda$ et $b = \log K + \log n$ et donc directement une estimation de K et λ.

Le problème est que cette estimation des paramètres de la distribution de Gumbel par Monte-Carlo ne permet pas de s'affranchir des hypothèses sous lesquelles l'approximation

de Gumbel est valide, à savoir l'indépedance des variables aléatoires ($\mathcal{S}_i$) ainsi qu'une large taille de séquence (n). De plus la précision de l'estimation des paramètres dépend du nombre de simulations réalisées. Par conséquent il n'y a pas vraiment intérêt à utiliser cette approche par rapport à une estimation directe de la significativité du Score Local par Monte-Carlo.

Complexité de l'Algorithme 4

Dans le pire des cas, cet algorithme possède une complexité de l'ordre de n^2 ou n est la taille de la séquence. Ce cas correspond à la situation où les régions détectées sont à chaque fois de taille 1. On montre ci-dessous que cet algorithme possède en pratique, une complexité moyenne de l'ordre de $n \log n$ dans le cas où les éléments de la séquence $\mathbb{S}$ sont indépendants et identiquement distribués et d'espérance négative.

Le fait que l'espérance soit négative nous permet de négliger la longueur de la région détectée par rapport à la longueur de la séquence (n). Une itération de l'algorithme va donc couper la séquence en deux parties. La question de la complexité de l'algorithme se ramène en ces termes : combien va-t-il falloir d'itérations pour n'avoir plus que des séquences de taille 1 ? La réponse est k tel que $n = 2^k$, en supposant que la séquence soit d'une taille qui soit une puissance de 2. On en déduit que $k = \log_2(n)$, où $\log_2$ est le logarithme en base 2. Étant donné d'autre part que la recherche de la région de Score Local maximal est de complexité $O(n_i)$ pour une sous-séquence de taille n_i, la complexité de la recherche à une étape donnée vaut $\sum O(n_i) = O(\sum n_i)$. Finalement, la complexité globale moyenne vaut : $k \times O(n)$ soit $O(n \log N)$.

Bibliographie

Agresti, A. (2002). *Categorical data analysis*. Wiley.

Ahrens, W. and Pigeot, I. (2004). *Handbook of epidemiologie*. Springer.

Allison, D., Gadbury, G., Heo, M., and et al (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, **39**, 1–20.

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipamn, D. (1990). Basic local alignement search tool. *Journal of Molecular Biology*, **215**, 403–410.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386.

Aschard, H., Guedj, M., and Demenais, F. (2007). A two-step multiple-marker strategy for genome-wide association studies. *BMC Genetics [in press]*.

Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**, 781–791.

Barnard, G. A. (1945). A new test for 2x2 contingency tables. *Nature*, **156**, 177.

Bates, J. and Constable, R. (1985). Proofs as programs. *ACM Transactions on Programming Languages and Systems*, **7**, 113–136.

Bellman, R. (1961). *Adaptive control processes*. Princeton University Press.

Benjamini, T. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple-testing. *Journal of the Royal Statistics Society B*, **57**, 289–300.

Billingsley, R. (1968). *Convergence of probability measures*. Wiley.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes : past successes for mendelian disease, future approaches for complex diseases. *Nature Genetics*, **33**, 228–237.

Bourgain, C., Génin, E., Cox, N., and Clerget-Darpoux, F. (2006). Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases. *European Journal of Human Genetics*, **15**, 260–263.

Brendel, V., Bucher, P., Nourbakhsh, I., Blaisdell, B., and Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *PNAS*, **89**, 2002–2006.

Breslow, N. and Day, N. (1982). *Statistical methods in cancer resarch*. Schlesselman.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K., Hayward, B., Keith, T., and Van Eedewegh, P. (2005). Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, **28**, 171–182.

Campbell, H. and Rudan, I. (2002). Interpretation of genetic association studies in complex disease. *Pharmacogenomics*, **2**, 349–360.

Cargill, M. and et al (1999). Characterization of snps in coding regions of human genes. *Nature Genetics*, **22**, 231–238.

Chapman, J., Cooper, J., Todd, J., and Clayton, D. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags : a class of tests and the determinants of statistical power. *Human Heridity*, **56**, 18–31.

Chumakov, I. and et al (2002). Genetic and physiological data implicating the new human gene g72 and the gene daao in schizophrenia. *PNAS*, **99**, 13675–13680.

Cochran, W. G. (1954). Some methods of strengthening the common chi-square tests. *Biometrics*, **10**, 417–451.

Collins, A., Lonjou, C., and Morton, N. (1999). Genetic epidemiology of snps. *PNAS*, **96**, 15173–15177.

Collins, A., Ennis, S., Taillon-Miller, P., Kwok, P. Y., and Morton, N. (2001). Allelic association with snps : metrics, populations and the linkage disequilibrium map. *Human Mutation*, **17**, 255–262.

Coulonges, C., Delaneau, O., Girard, M., Do, H., Adkins, R., Spadoni, and Zagury, J.-F. (2006). Computation of haplotypes on snps subsets : advantage of the global method. *BMC Genetics*, **7**, 50.

Cox, D. and Kraft, P. (2006). Quantification of the power of hardy-weinberg equilibrium testing to detect genotyping errors. *Human Heredity*, **61**, 10–14.

Dalmasso, C., Bar-Hen, A., and Broët, P. (2007). A constrained polynomial regression procedure for estimating the local false discovery rate. *submitted*.

Daly, M., Rioux, J., Schaffner, S., Hudson, T., and Lander, E. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–232.

Davies, R. (1973). Numerical inversion of characteristic function. *Biometrika*, **60**, 481–482.

Davies, R. (1980). The distribution of a linear combination of chi-square random vriables. *Applied Statistics*, **29**, 323–333.

Dempster, A., Laird, N., and Rubin, D. (197). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.

Deng, H. (2001). Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. *Genetics*, **159**, 1319–1323.

Detera-Wadleigh, S. and McMahon, F. (2006). G72/g30 in schizophrenia and bipolar disorder : review and meta-analysis. *Biological Psychiatry*, **60**, 106–114.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Efron, B. (2004). Large-scale simultaneous hypothesis testing : the choice of a null hypothesis. *Journal of American Statistical Association*, **99**, 96–104.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical bayes analysis of microarray experiment. *Journal of American Statistical Association*, **96**, 1151–1160.

Egan, S., Nathan, P., and Lumley, M. (2003). Diagnostic concordance of icd-10 personality and comorbid disorders : a comparison of standard clinical assessment and structured interviews in clinical setting. *Psychiatry*, **37**, 484–491.

Elston, R., Olson, J., and Palmer, L., editors (2002). *Biostatistical Genetics anf Genetic Epidemiology*. Wiley and sons.

Ewen, K., Bahlo, M., Treloar, S., Levinson, D., Mowry, B., J, B., and Foote, J. (2000). Identification and analysis of errors types in high-througput genotyping. *American Journal of Human Genetics*, **67**, 727–736.

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

Feder, J., Gnirke, A., Thomas, W., Tsuchihashi, Z., and et al (1996). A novel mhc class i-like gene is mutated in patients with hereditary hameochromatosis. *Nature Genetics*, **13**, 399–408.

Fisher, R. A. (1934). *Statistical methods for research workers*. Oliver and Boyd.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. Wiley.

Forner, K., Lamarine, M., Guedj, M., Dauvillier, J., and Wojcik, J. (2007). Universal false discovery rate estimation methodology for genome-wide association studies. *submitted*.

Forster, J., McDonald, J., and Smith, P. (1996). Monte-carlo exact conditional tests for log-linear and logisitc models. *Journal of the Royal Statistics Society B*, **58**, 445–453.

Gabriel, S. and et al (2002a). Segregation at three loci explains familial and population risk in hirschsprung disease. *Nature Genetics*, **31**, 89–93.

Gabriel, S. and et al (2002b). The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Garnier, S. (2007). *Recherche des facteurs génétiques de susceptibilité à la Polyarthrite Rhumatoïde*. Ph.D. thesis, Université d'Evry Val d'Essone.

Goldstein, D. B. (2001). Island of linkage disequilibrium. *Nature Genetics*, **29**, 109–111.

Gordon, D., Finch, S., Nothnagel, M., and Ott, J. (2002). Power and sample size calcultation for case-control genetic association tests when errors are present : application to snps. *Human Heredity*, **54**, 22–33.

Green, P., Lid-Hjort, N., and Richardson, S. (2003). *Hightly structured stochastic systems*. Oxford University Press.

Guedj, M. (2007). Statistical interpretation of hardy-weinberg disequilibrium in cas-control association studies. *submitted*.

Guedj, M., Della-Chiesa, E., Picard, F., and Nuel, G. (2006a). Computing power in case-control association studies through use of quadratic approximations : application to meta-statistics. *Annals of Human Genetics*, **71**, 262–270.

Guedj, M., Robelin, D., Hoebeke, M., Lamarine, M., Wojcik, J., and Nuel, G. (2006b). Detecting local high-scoring segments : a first stage approach to genome-wide association studies. *Statistical Applications to Genetic and Molecular Biology*, **5**, 22.

Guedj, M., Wojcik, J., Della-Chiesa, E., Nuel, G., and Forner, K. (2006c). A fast, unbiased and exact allelic test for case-control association studies. *Human Heredity*, **61**, 210–221.

Hall, P. (1981). On the nonparametric estimation of mixture proportions. *Royal Statistics Society B*, **43**, 147–156.

Hanson, R., Craig, D., Millis, M., Yeatts, K., Kobes, S., Pearson, J., Lee, A., Knowler, W., Nelson, R., and Wolford, J. (2007). Indentification of pvt1 as a candidate gene for end-stage renal disease in type 2 diabetes using a pooling-based genome-wide snp association study. *Diabetes*, **56**, 975–983.

Hao, K., Xu, X., Laird, N., Wang, X., and Xu, X. (2004). Power estimation of multiple snp association test of case-control study and application. *Genetic Epidemiology*, **26**.

Heidema, A., Boer, J., Nagelkerke, N., Marimam, E., van der A, D., and Feskens, E. (2006). The challenge for genetic epidemiologists : how to analyze large numbers of snps in relation to complex diseases. *BMC Genetics*, **7**, 23.

Hendel, H., Cho, Y., Gauthier, N., Rappaport, J., Schachter, F., and Zagury, J. (1996). Contribution of cohort studies in understanding hiv pathogenesis : introduction of the griv cohort. *Biomedical Pharmacotherapy*, **50**, 480–487.

Herbert, A., Gerry, N., McQueen, M., and et al (2006). A common genetic variant is associated with adult and childhood obesity. *Science*, **312**, 279–283.

Hirschhorn, J. and Daly, M. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Review Genetics*, **6**, 95–108.

Hogben, L. (1932). The genetic analysis of familial traits. ii. double gene substitutions, with special reference to hereditary dwarfism. *Journal of Genetics*, **25**, 211–240.

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Review Genetics*, **4**, 701–709.

Hoh, J., Wille, A., and Ott, J. (2001). Trimming, weighting and grouping snps in human case-control association studies. *Genome Research*, **11**, 2115–2119.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., and Xu, C. (2004). Detection of genotyping errors by hardy-weinberg equilibrium. *European Journal of Human Genetics*, **12**, 395–399.

Ibrahim, S. and Gold, R. (2005). Genomics, proteomics, metabolomics : what is in a word for multiple-sclerosis. *Current opinion in Neurology*, **18**, 231–235.

Iglehart, D. (1972). Extreme values in the gi/g/1 queues. *Annals of Mathematical Statistics*, **43**, 627–635.

Ioannidis, J. (2003). Genetic associations : false or true ? *Trends in Molecular Medecine*, **9**, 135–138.

Ioannidis, J., Ntzani, E., Trikalinos, T., and Contopoulos-Ioannidis, D. (2001). Replication validity of genetic association studies. *Nature Genetics*, **29**, 306–309.

Jackson, M., Genin, E., Knapp, M., and Escary, J. (2002). Acurate power aproximations for chi-square tests in case-control association studies of complex disease genes. *Annals of Human Genetics*, **66**, 307–321.

Jannot, A. (2004). *Détection et modélisation de facteurs de risques génétiques dans des maladies multifactorielles*. Ph.D. thesis, Université Paris 11.

Ji, F., Yang, Y., Haynes, C., Finch, S., and Gordon, D. (2005). Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors. *Statistics Applied to Genetic and Molecular Biology*, **4**.

Jorde, L. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Research*, **10**, 1435–1444.

Kang, S., Gordon, D., and Finch, S. (2004). What snps genotyping errors are most costly for genetic association studies. *Genetic Epidemiology*, **26**, 132–141.

Karlin, S. (2005). Statistical signals in bioinformatics. *PNAS*, **102**, 13355–13362.

Karlin, S. and Altschul, S. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *PNAS*, **90**, 5873–5877.

Karlin, S. and Brendel, V. (1992). Chance and significance in protein and dna sequences analysis. *Science*, **257**, 39–49.

Karlin, S. and Dembo, A. (1992). Limit distributions of maximal segmental score among markov-dependent partial sums. *Advances in Applied Probability*, **24**, 113–140.

Karlin, S., Bucher, P., Brendel, V., and Altschul, S. (1991). Statistical methods for insights for protein and dna sequences. *Annual Review of Biophysics and Biophysical Chemistry*, **20**, 175–203.

Keightley, P. and Knott, S. (1999). Testing the correspondance between map positions of quantitative trait loci. *Genetical Research*, **74**, 332–338.

Kerem, B. and et al (1989). Identification of the cystic fibrosis gene : genetic analysis. *Science*, **245**, 1073–1080.

Knapp, M. (2003). Re : baised tests of association : comparison of alleles frequencies when departing from hardy-weinberg proportions. *American Journal of Epidemiology*, **153**, 287.

Knowler, W., Williams, R., Pettit, D., and Steinberg, A. (1988). Gm3 ;5 ;13 ;14 and type 2 diabetes mellitus : an association in american indians with genetic admixture. *American Journal of Human Genetics*, **43**, 520–526.

Koehler, K. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistics Association*, **75**, 336–344.

Kruglyak, L. (1999). Prospect for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**, 139–144.

Lamy, P., Andersen, C., Wikman, F., and Wiuf, C. (2006). Genotyping and annotation of affymetrix snps array. *Nucleic Acids Research*, **34**, 14.

Lander, E. and Kryglyak, L. (1995). Genetic dissection of complex traits : guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241–247.

Lantowski, K. and Makalowski, W. (2000). Methodological function of hypotheses in science : old ideas in new cloth. *Genome Research*, **10**, 273–274.

Leal, S. (2005). Detection of genotyping errors and pseudo-snps via deviations from hardy-weinberg equilibrium. *Genetic Epidemiology*, **29**, 204–214.

Leroux, B. (1992). Consistent estimation of mixing distribution. *Annals of Statistics*, **20**, 1350–1360.

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

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data*. Wiley and sons.

Liu, Y., Liu, Y., Recker, R., and Deng, H. (2003). Molecular studies of identification of genes for osteoporosis : the 2002 update. *Journal of Endocrinology*, **177**, 147–196.

Longmate, J. (2001). Complexity and power in case-control association studies. *American Journal of Human Genetics*, **68**, 1229–1237.

Lu, Z. and King, M. (2002). Improving the numerical technique for computing the accumulated distribution of a quadratic form in normal variables. *Econometric Reviews*, **21**, 149–165.

Lunetta, K., Hayward, L., Segal, J., and Van Eerdewegh, P. (2004). Screening large-scale association study data : exploiting interactions using random forest. *BMC Genetics*, **5**, 32.

Maclure, M. and Schneeweiss, S. (2001). Causation of bias : the episcope. *Epidemiology*, **12**, 114–122.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from restropective studies of disease. *Journal of the National Cancer Institute*, **22**, 719–748.

Maraganore, D., de Andrade, M., Lesnick, T., and et al (2005). High-resolution whole-genome association study of parkinson disease. *American Journal of Human Genetics*, **77**, 685–693.

Martin, M. e. a. (2002). Epistatic interaction between kir3ds1 and hla-b delays the progression to aids. *Nature Genetics*, **31**, 429–434.

Matsuzaki, H., Dong, S., Loi, H., Di, X., and Liu, G. (2004). Genotyping over 100,000 snps on a pair of oligonucleotide arrays. *Nature Methods*, **1**, 109–111.

McGullagh, P. and Nedler, J. (1989). *Generalized linear models*. Chapman and Hall.

McLachlan, G. and Krishnan, T. (1996). *The EM algorithm and extensions*. John Wiley and Sons, New York.

McLachlan, G. and Krishnan, T. (2000). *Finite Mixture Models*. John Wiley and Sons, New York.

McLachlan, G., Bean, R., and Ben-Tovin Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**, 1608–1615.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 651–655.

Mercier, S. (1999). *Statistiques des scores pour l'analyse et la comparaison de séquences biologiques*. Ph.D. thesis, Université de Rouen.

Mitra, S. (1958). On the limiting power function of the frequency chi-square test. *Annals of Mathematical Statistics*, **29**, 1221–1233.

Moore, J. and Williams, S. (2002). New strategies for identifying gene-gene interactions in hypertension. *Annals of Medecine*, **34**, 88–95.

Morris, A., Whittaker, J., and Balding, D. (2003). Multi-point linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity. *PNAS*, **11**, 13442–13446.

Morton, N. (1998). Significance levels in complex inheritance. *American Journal of Human Genetics*, **62**, 690–697.

Mourant, A., Kopec, A., and Domansiewska-Sobczak, K. (1976). *The distribution of the human blood groups and other polymorphisms*. Oxford University Press.

Nelson, M., Kardia, S., Ferell, R., and Sing, C. (2001). A combinatorial partitioning method to identifiy multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, **11**, 458–470.

Newton-Cheh, C. and Hirschhorn, J. (2005). Genetic association studies of complex traits : design and analysis issues. *Mutation Research*, **573**, 54–59.

Nielsen, D., Ehm, M., and Weir, B. (1999). Detecting marker-disease association by testing for hardy-weinberg disequilibrium at a marker locus. *American Journal of Human Genetics*, **63**, 1531–1540.

Nielsen, D., Ehm, M., Zaykin, D., and Weir, B. (2004). Effect of two and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics*, **168**, 1029–1040.

Nordborg, M. and Tavare, S. (2002). Linkage disequilibrium : what history has to tell us. *Trends in Genetics*, **18**, 83–90.

Page, G., Varghese, G., Go, R., Page, P., and Allison, D. (2003). Are we there yet ? deciding when one has demonstrated specific genetic causation in complex diseases and quantative traits. *American Journal of Human Genetics*, **73**, 711–719.

Parsons, C., Mroczkowski, H., McGuigan, F., Albagha, O., Manolagas, S., Reid, D., Ralston, S., and Reis, R. (2005). Interspecies synteny mapping identifies a quantitative trait locus for bone mineral density on human chromosome xp22. *Human Molecular Genetics*, **14**, 3141–3148.

Peduzzi, P., Concato, J., Kemper, E., Holford, T., and Feinstein, A. (1996). A simulation study of the number of events per variable in logisitc regression analysis. *Journal of Clinical Epidemiology*, **49**, 1373–1379.

Peltonem, L. (2000). Positional cloning of disease genes : advantages of genetic isolates. *Human Heredity*, **50**, 66–75.

Price, A. and et al (2006). Principal component analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

Pritchard, J., Stephens, M., Rosenberg, N., and Donnelly, P. (2000). Association mapping in structured populations. *American Journal of Human Genetics*, **67**, 170–181.

Province, M., Shannon, W., and Rao, D. (2001). Classification methods for confronting heterogeneity. *Advanced Genetics*, **42**, 273–286.

Race, R. and Sanger, R. (1975). *Blood groups in man*. Blackwell.

Reif, D., White, B., and Moore, J. (2004). Integrated analysis of genetic, genomic and proteomic data. *Expert Review in Proteomics*, **1**, 67–75.

Rice, J., Saccone, N., and E, R. (2001). *Definition of the phenotype in Genetic dissection of complex traits*. Academic Press, San Diego.

Riley, B. and Kendler, K. (2006). Molecular genetics studies of schizophrenia. *European Journal of Human Genetics*, **14**, 669–680.

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

Risch, N. and Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human dieases i. dna poling. *Genome Research*, **8**, 1273–1288.

Ritchie, M., Hahn, L., Rood, N., Bailey, R., Dupont, W., Parl, F., and Moore, J. (2001). Mutifactor-dimensionalty reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, **69**, 138–147.

Ritchie, M., Lance, W., and Moore, J. (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy and genetic heterogeneity. *Genetic Epidemiology*, **24**, 150–157.

Robelin, D. (2005). *Détection de court segments inversés dans les génomes : méthodes et applications*. Ph.D. thesis, Université Paris 5.

Robin, S., Bar-Hen, A., Daudin, J., and Pierre, L. (2007). A semi-parametric approach for mixture models : application to local false discovery rate estimation. *Computational Statistics and Data Analysis (to appear)*.

Roychoudhuri, A. K. and Nei, M. (1988). *Human polymorphic genes word distribution*. Oxford University Press.

Rubin, D. (1987). *Multiple Imputation for nonresponse in surveys*. Wiley and sons.

Ruzzo, W. and Tompa, M. (1999). A linear time algorithm for finding all maximal scoring subsequences. In *7th Int. Conf. Intelligent Systems for Molecular Biology*, pages 234–241.

Sasieni, P. (1997). From genotype to genes : doubling the sample size. *Biometrics*, **53**, 1253–1261.

Satten, G., Flanders, W., and Yang, Q. (2001). Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics*, **68**, 466–477.

Schaid, D. (2004). Evaluating association of haplotypes with traits. *Genetic Epidemiology*, **27**, 348–364.

Schaid, D. and Jacobsen, S. (1999). Biased tests of association : comparisons of allele frequencies when departing from hardy-weinberg proportions. *American Journal of Epidemiology*, **149**, 706–711.

Setakis, E., Stirnadel, H., and Balding, D. (2006). Logistic regression protects against population structure in genetic association studies. *Genome Research*, **16**, 290–296.

Sham, P., Cherny, S., Purcell, S., and Hewitt, J. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance component models for sibship data. *American Journal of Human Genetics*, **66**, 1616–1630.

Shen, H., Liu, Y., Liu, P., Recker, R., and Deng, H. (2005). Nonreplication in genetic studies of complex diseases : lessons learned form studies of osteoporosis and tentative remedies. *Journal of Bone and mineral research*, **20**, 325–376.

Sillanpaa, M. and Auranen, K. (2004). Replication in genetic studies of complex traits. *Annals of Human Genetics*, **68**, 646–657.

Slager, S. and Schaid, D. (2001). Case-control studies of genetic markers : power and sample size approximations for armitage's test for trend. *Human Heredity*, **52**, 149–153.

Song, K. and Elston, R. (2006). A powerful method of combining measures of association and hardy-weinberg disequilibrium for fine-mapping in case-control studies. *Statistics in Medecine*, **25**, 105–126.

Song, K., Orloff, M., Lu, Q., and Elston, R. (2005). Fine-mapping using the weighted average method for a case-control study. *BMC Genetics, Proceedings of GAW14* , **6**.

Spielman, R. and W, E. (1998). A sibship test for linkage disequilibrium in the presence of association : the sib transmission/disequilibrium test. *American Journal of Human Genetics*, **62**, 450–458.

Spielman, R., McGinnis, R., and W, E. (1993). Transmission test for linkage disequilibrium : the insulin gene region and insulin dependent diabetes mellitus. *American Journal of Human Genetics*, **52**, 450–458.

Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Storey, D. and R, T. (2003). Statistical significance for genomewide studies. *PNAS*, **100**, 9440–9445.

The-International-HapMap-Consortium (2004). Integrating ethics and science in the international hapmap project. *Nature Review Genetics*, **5**, 467–475.

Thomas, A. and Camp, N. (2004). Graphical modeling of the joint distribution of alleles at associated loci. *American Journal of Human Genetics*, **74**, 1088–1101.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, **58**, 267–288.

Traherne, J., Barcellos, L., Sawcer, S., Compston, A., Ramsay, P., Hauser, S., Oksenberg, J., and Trowsdale, J. (2006). Association of the truncating splice site mutation in btnl2 with multiple sclerosis is secondary to hla-drb1*15. *Human Molecular Genetics*, **15**, 155–161.

Tregouet, D., Escolano, S., Tiret, L., Mallet, A., and Gollmard, J. (2004). A new algorithm for haplotype-based association analysis : the stochastic-em algorithm. *Annals of Human Genetics*, **68**, 165–177.

Trevor-Roper, P. (1952). Marriage of two complete albinos with normally pigmented offsrpring. *British Journal of Ophthalmology*, **36**, 107–110.

Tu, I. and Whittemore, A. (1999). Power of association and linkage tests when the disease alleles are unobserved. *American Journal of Human Genetics*, **64**, 641–649.

Wacholer, S., Rothman, N., and Caporaso, N. (2000). Population stratification in epidemiologic studies of common genetic variants and cancer : quantification of bias. *Journal of the National Cancer Institute*, **92**, 1151–158.

Wald, A. (1943). Test of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. American Mathematical Society*, **54**, 426–482.

Wedderburn, R. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63**, 27–32.

Weinberg, C. and Morris, R. (2003). Invited commentary : testing for hardy-weinberg disequilibrium using a genome single-nucleotid polymorphism scan based on cases only. *American Journal of Epidemiology*, **158**, 401–403.

Whittemore, A. and Halpern, J. (2001). Problems in the definition, interpretation and evalutation of genetic heterogeneity. *American Journal of Human Genetics*, **68**, 457–465.

Willison, D., Keshavjee, K., Nair, K., Goldsmith, C., and Holbrook, A. (2003). Patient's consent preferences for research uses of information in electronic medical records : interview and survey data. *Br Medical Journal*, **326**, 373–378.

Wright, S. (1921). Systems of mating. *Genetics*, **6**, 111–178.

Xiong, M., Feghali-Bostwick, C., Arnett, F., and Zhou, X. (2005). A systems biology approach to genetic studies of complex diseases. *Federation of European Biochemical Societies Letters*, **579**, 5325–5332.

Yate, F. (1934). Contingency tables involving small numbers and the chi-square test. *Journal of the Royal Statistics Society*, **1**, 217–235.

Yu, J. and et al (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203–208.

Zavattari, P., Deidda, E., Whalen, M., and et al (2000). Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations : demography, chromosome recombination frequency and selection. *Human Molecular Biology*, **9**, 2947–2957.

Zhang, K., Deng, M., Chen, T., Waterman, M., and Sun, F. (2002). A dynamic programming algortihm for haplotype block partitioning. *PNAS*, **99**, 7335–7339.

Zondervan, K. and Cardon, L. (2004). The complex interplay among factors that influence allelic association. *Nature Review Genetics*, **5**, 89–100.