N° d'ordre : ? ? ?
N° bibliothèque : ? ? ?

Université Paris VI
Unité INSERM U535

THÈSE

pour obtenir le grade de

Docteur de l'Université Paris VI
Spécialité : Biologie

au titre de l'école doctorale de Diversité du Vivant

présentée et soutenue publiquement le 12/12/2005

par Claire BARDEL

Mise en évidence de facteurs génétiques de risque en utilisant des phylogénies d'haplotypes

Devant la commission d'examen formée de :

Monsieur	Pierre	DARLU,	Directeur de thèse
Madame	Émanuelle	GÉNIN,	Directrice de thèse
Madame	Heike	BICKEBOLLER,	Rapporteur
Monsieur	Manolo	GOUY,	Rapporteur
Monsieur	Laurent	GOUYA,	Examinateur
Madame	Évelyne	HEYER,	Examinatrice
Monsieur	Hervé	LE GUYADER,	Examinateur
Monsieur	Hadi	QUESNEVILLE,	Examinateur

Remerciements

Résumé : Avec le développement des techniques de biologie moléculaire, de plus en plus de marqueurs génétiques peuvent être analysés pour rechercher des facteurs génétiques de risque pour des maladies. L'utilisation d'haplotypes permet de prendre en compte l'information conjointe de plusieurs marqueurs. Mais, avec l'augmentation du nombre de marqueurs disponibles, le nombre d'haplotypes possibles devient très élevé, ce qui rend la puissance des tests haplotypiques d'autant plus faible. Afin d'améliorer la puissance des ces tests, une stratégie consiste à regrouper les haplotypes et à effectuer les tests sur ces groupes. Au cours de ce travail, nous nous sommes intéressés à une type très particulier de stratégie de regroupement : les méthodes basées sur les phylogénies. Elles consistent à utiliser l'arbre représentant l'histoire évolutive des haplotypes pour identifier des groupes d'haplotypes sur-représentés chez les malades. Les mutations à l'origine de ces groupes sont potentiellement impliquées dans la susceptibilité à la maladie.

Dans un premier temps, nous avons mis au point une nouvelle méthode utilisant des phylogénies pour détecter une association entre un gène et une maladie et pour localiser finement les marqueurs impliqués dans son déterminisme. Nous l'avons implémentée dans le logiciel `ALTree` . L'efficacité d'`ALTree` a été testée par simulations et comparée à celle d'autres tests. Nous avons montré que l'utilisation de phylogénies n'entraîne pas d'importants gains de puissance pour la détection de l'association. En revanche, cela permet d'améliorer l'efficacité d'identification des sites de susceptibilité dans le cas où plusieurs locus en interaction sont impliqués dans le déterminisme de la maladie. Nous nous sommes aussi intéressés à l'influence de différents facteurs sur `ALTree` . Nous avons montré que son efficacité dépend beaucoup des paramètres du modèle génétique simulé pour la maladie (pénétrances et fréquences des allèles à risque) et des conditions de reconstruction de l'arbre phylogénétique (méthode de reconstruction de l'arbre, méthode d'optimisation des caractères, enracinement).

Nous avons ensuite étudié la puissance d'une autre méthode basée sur des phylogénies : `CLADHC`. Elle permet l'analyse de longues séquences car elle reconstruit l'arbre et elle réalise l'analyse statistique dans une fenêtre qui se déplace le long des séquences de marqueurs. Ses auteurs ont montré qu'elle permet un très fort gain de puissance pour le test d'association. Nos propres simulations ont, en revanche, souligné que cette méthode n'est généralement pas plus puissante qu'un test classique d'association allélique.

Enfin, nous avons appliqué notre méthode `ALTree` à des données. Elle nous a permis de retrouver une association forte entre un gène et une maladie sur des données simulées pour un atelier de travail international. Nous avons aussi pu confirmer l'implication de 3 variants du gène CARD15 dans la susceptibilité à la maladie de Crohn et d'un variant du gène DRD2 dans la susceptibilité à la schizophrénie. Ces résultats montrent la validité des méthodes basées sur des phylogénies pour la recherche de facteurs génétiques de risque.

Mots-clés : phylogénie, études d'association, maladies mutifactorielles, haplotypes, localisation de sites de susceptibilité

Abstract: With the developpement of molecular biology, more and more markers are now available within genes. These makers can be used to look for genetic risk factors for complex diseases. To take into account the joint information on several markers, haplotypic methods can be used. But, with the increasing number of markers available, the number of haplotypes is now large and thus, the power of haplotypic tests may be low. To increase the power of these tests, a strategy consists in grouping haplotypes and in performing the tests on these groups. In this work, we focus on phylogeny-based grouping methods. They consist in reconstructing a tree corresponding to the evolutionary history of haplotypes and to use it to identify haplotype groups containing significantly more cases than controls. Mutations defining these groups are putative susceptibility sites.

We have first developed a new phylogeny-based method to look for association between a candidate gene and a disease and to identify the variants involved in the determinism of the disease. This method is implemented in the `ALTree` software. Simulations were performed to evaluate its efficiency and to compare it to other methods' efficiency. We show that the use of phylogenies does not improve the power to detect association. However, it can increase the efficiency to localize the susceptibility sites especially when several interacting loci are involved in the determinism of the disease. We also study the influence of several factors on `ALTree` 's performances. We find that its efficiency highly depends on the disease model (penetrances and susceptibility allele frequencies) and on the phylogenetic tree reconstruction parameters (reconstruction method, character states optimization method, rooting method).

Then, we have evaluated the power of another phylogeny-based method: `CLADHC`. It allows to deal with very long marker sequences because the phylogenetic tree reconstruction and the statistical analysis are performed in overlapping windows, sliding along the sequence. According to its authors, `CLADHC` can greatly increase the power of an association test. However, our own simulations show that, in general, it is no nore powerful than a simple allelic association test.

`ALTree` has finally been applied to three data sets. We have been able to detect a strong association between a gene and a disease on date simulated for the 12th Genetic Analysis Workshop. We have also confirmed the role of three variants in the CARD15 gene in the susceptibility to Crohn disease and the role of a variant in the DRD2 gene, involved in the susceptibility to schizophrenia. These results show the validity of phylogeny-based methods to look for disease susceptibility loci.

Keywords: phylogeny, association studies, complex diseases, haplotypes, localisation of susceptibility sites

Sommaire

Introduction **1**

1 Introduction à la génétique épidémiologique et à la phylogénie **3**
1.1 La génétique épidémiologique . . . 3
1.2 La phylogénie . . . 23

2 L'utilisation des phylogénies d'haplotype en génétique épidémiologique **35**
2.1 L'analyse des clades emboîtés (Templeton *et al.*, 1987) . . . 37
2.2 La méthode `EHAP` (Seltman *et al.*, 2003) . . . 41
2.3 La méthode `CLADHC` (Durrant *et al.*, 2004) . . . 44
2.4 La méthode de Tree Scanning (Templeton *et al.*, 2005) . . . 49
2.5 La méthode de Tzeng (Tzeng, 2005) . . . 51
2.6 Utilisation des méthodes phylogénétiques sur des données . . . 53
2.7 Bilan sur les méthodes d'association et de localisation de sites de susceptibilité basées sur des phylogénies . . . 59

3 La méthode `ALTree` **61**
3.1 Principe de la méthode `ALTree` . . . 62
3.2 Implémentation de la méthode au sein du logiciel `ALTree` . . . 67
3.3 Étude de puissance . . . 71
3.4 Les facteurs influençant la méthode . . . 91
3.5 Étude de puissance d'une méthode phylogénétique alternative . . . 104

4 Application de la méthode à des données **121**
4.1 Les données simulées pour GAW12 . . . 122
4.2 La maladie de Crohn . . . 128
4.3 La schizophrénie . . . 133

Conclusion **139**

A Première méthode de localisation du site de susceptibilité **143**

B Position des marqueurs et fréquence de leurs allèles dans les différentes régions du génome étudiées **145**
B.1 Données de la base Seattles SNPs . . . 145
B.2 Données de la base Hapmap . . . 146

Bibliographie **147**

Liste des publications **157**

Index **159**

Index des auteurs cités **163**

Table des figures **167**

Liste des tableaux **169**

Table des matières **171**

Introduction

L'identification de facteurs génétiques de risque pour les maladies est un des enjeux majeurs en génétique humaine. De nombreuses recherches portent donc sur le développement de méthodes d'analyses du déterminisme génétique des maladies.

Historiquement, les premiers facteurs de risque identifiés correspondaient à des maladies rares, à forte concentration familiale. Les généticiens ont montré que la plupart de ces maladies sont dues à une mutation dans un seul gène. Ces maladies ont été qualifiées de *monogéniques*.

Plus récemment, on s'est intéressé à un second type de maladies dites maladies *multifactorielles*. Il s'agit de maladies relativement fréquentes telles que les cancers, le diabète, les maladies psychiatriques ou les maladies cardio-vasculaires. Le déterminisme de ces maladies est encore mal connu, mais il serait dû à des facteurs génétiques et à des facteurs environnementaux dont les effets peuvent éventuellement se combiner. La nature des facteurs et leur part respective dans la pathologie varie d'une maladie à l'autre. Les caractéristiques des facteurs génétiques de risque des maladies multifactorielles sont différentes de celles des maladies monogéniques. Ce ne sont pas des mutations délétères rares, mais plutôt des variants, généralement fréquents, dont la présence n'est ni nécessaire ni suffisante au développement de la maladie chez un individu. C'est pourquoi on ne parle pas de mutations, mais uniquement de facteurs de risque : ces variants confèrent des risques de développer la maladie qui varient en fonction des génotypes des individus qui les portent.

Les premières méthodes permettant d'identifier et de localiser les facteurs génétiques de risque ont été développées dans le cadre de l'étude des maladies monogéniques. Elles utilisent l'information apportée par des marqueurs génétiques, c'est-à-dire des séquences d'ADN polymorphes dont la localisation sur le génome est généralement connue. Elles ont permis d'identifier avec succès de nombreux gènes impliqués dans le déterminisme de maladies monogéniques mais, lorsqu'elles ont été appliquées à des maladies multifactorielles, les résultats ont été beaucoup moins probants : plusieurs études ont mis en évidence des facteurs génétiques de risque, mais ces résultats ont rarement été répliqués. Les méthodes développées dans le cadre des maladies monogéniques ne sont donc pas adaptées à l'étude des maladies multifactorielles et de nouvelles approches ont dû être développées.

Dans les quinze dernières années, le développement des techniques de biologie moléculaire a été très important. Le nombre de marqueurs connus sur le génome a littéralement explosé, permettant la mise au point de cartes de marqueurs très denses. Les différents marqueurs peuvent être utilisés individuellement ou de façon conjointe, sous la forme d'haplotypes. Cette dernière stratégie peut être particulièrement efficace dans le cas des mala-

dies multifactorielles car elle permet de prendre en compte d'éventuelles interactions entre plusieurs facteurs de risque situés sur le même chromosome. De nombreuses méthodes haplotypiques ont alors été développées, mais, avec l'augmentation du nombre de marqueurs disponibles, le nombre d'haplotypes différents augmente aussi, ce qui limite la puissance de ces méthodes. De plus, lorsque le nombre d'haplotypes est élevé, le nombre d'individus portant chaque haplotype risque d'être faible, ce qui pose des problèmes de petits effectifs pour les tests statistiques.

Afin de pallier ces problèmes, différentes méthodes de regroupement des haplotypes ont été proposées. L'une des stratégies possibles consiste à utiliser l'information apportée par la reconstruction de l'histoire évolutive des haplotypes : le regroupement se fait alors en fonction de la proximité évolutive des haplotypes. En comparant les différents groupes d'haplotypes entre eux, cette approche permet de tester si un gène particulier est impliqué dans le déterminisme d'une maladie.

Au cours de cette thèse, nous avons cherché à évaluer l'intérêt que peut présenter l'utilisation de l'histoire évolutive des haplotypes pour la recherche de facteurs génétiques de risques des maladies multifactorielles. Cette stratégie est-elle plus efficace que les stratégies classiques utilisées jusqu'à présent et dans quelles conditions ? Pour répondre à cette question, nous avons été conduits à développer une nouvelle méthode basée sur la reconstruction de l'histoire évolutive d'haplotypes que nous avons appelée `ALTree` . Elle permet non seulement de tester si un gène est impliqué dans le déterminisme d'une maladie mais aussi, de localiser précisément les facteurs de risque sur ce gène.

Ce travail se situe à l'interface entre deux domaines : la génétique épidémiologique, dont le but est l'identification des facteurs génétiques de risque, et la phylogénie dont le but est la reconstruction de l'histoire évolutive d'objets (ici, des haplotypes). Dans un premier temps, nous introduirons les concepts de base de ces deux domaines. Nous décrirons en particulier, les différentes approches utilisées pour étudier les maladies multifactorielles et les méthodes pour reconstruire l'histoire évolutive des marqueurs. Dans un second temps, nous nous intéresserons aux différentes méthodes d'identification de facteurs génétiques de risque basées sur l'utilisation de phylogénies développées jusqu'à présent et à leurs applications à des maladies. Nous introduirons ensuite la méthode `ALTree` . Nous étudierons sa puissance pour détecter si un gène est impliqué dans le déterminisme d'une maladie ainsi que son efficacité pour localiser le facteur de risque sur le gène dans différentes conditions (différents modèles de maladie, différentes conditions de reconstruction de l'histoire évolutive). En parallèle, nous étudierons la puissance d'une autre méthode d'identification de facteurs de risque basée sur des phylogénies : la méthode `CLADHC` (Durrant *et al.*, 2004). Enfin, dans une dernière partie, nous appliquerons la méthode `ALTree` à trois jeux de données, l'un simulé pour un atelier de travail international et les deux autres concernant respectivement la maladie de Crohn et la schizophrénie.

Chapitre 1

Introduction à la génétique épidémiologique et à la phylogénie

Sommaire

1.1 La génétique épidémiologique 3
1.1.1 Les concepts de base 4
1.1.1.1 Les polymorphismes 4
1.1.1.2 La notion de pénétrance 8
1.1.1.3 Les méthodes de recherche de facteurs de risque génétiques 9
1.1.2 Les méthodes d'association en stratégie gène candidat 11
1.1.2.1 Les différents tests d'association 11
1.1.2.2 Les méthodes de localisation des sites de susceptibilité . . . 16
1.1.2.3 Tests multiples et catégories rares 18
1.2 La phylogénie 23
1.2.1 Les concepts de base 23
1.2.1.1 Définition 23
1.2.1.2 Les arbres phylogénétiques 23
1.2.2 Les méthodes de reconstruction phylogénétique 25
1.2.2.1 Les méthodes de parcimonie 25
1.2.2.2 Les méthodes phénétiques 29
1.2.2.3 La méthode de maximum de vraisemblance 30
1.2.2.4 Les avantages et inconvénients de chaque méthode 33

1.1 La génétique épidémiologique

Un des enjeux majeurs des études en génétique épidémiologique est l'identification de facteurs génétiques de risque impliqués dans la susceptibilité à des maladies. Dans cette partie, nous allons tout d'abord décrire les objets utilisés en génétique épidémiologique — les polymorphismes génétiques — et nous allons brièvement introduire un des paramètres permettant de modéliser l'expression et la transmission des maladies : la pénétrance . Nous allons ensuite exposer les deux types d'approches utilisées actuellement pour rechercher des facteurs génétiques de risque, à savoir, les analyses de liaison et les analyses d'association. Enfin, nous nous intéresserons plus en détails aux méthodes d'association parmi lesquelles s'inscrit la méthode que nous avons mis au point au cours de cette thèse.

1.1.1 Les concepts de base

1.1.1.1 Les polymorphismes

On dit qu'un locus est polymorphe s'il existe sous au moins deux formes — ou allèles — dans la population. Le polymorphisme est le résultat de mutations de différents types : substitutions ou insertions/délétions. Du fait des progrès réalisés en biologie moléculaire, de plus en plus de locus polymorphes sont identifiés sur le génome.

Les différents types de polymorphismes

Les RFLPs Les RFLPs (*Restriction Fragment Length Polymorphisms*) sont des polymorphismes correspondant à des variations de taille des produits de la digestion de l'ADN génomique par les enzymes de restriction. Très souvent, ces variations de taille sont liées à la présence ou à l'absence d'un site de restriction, c'est-à-dire, d'une séquence d'ADN particulière qui est reconnue et coupée par une enzyme spécifique de cette séquence. Dans ce cas, le polymorphisme est biallélique (présence ou absence du site). Plus rarement, les variations de taille des fragments de restriction sont dus à des insertions ou des délétions dans les fragments. Ces polymorphismes RFLPs sont progressivement délaissés au profit des SNPs, plus nombreux et plus faciles à génotyper et à interpréter.

Les micro- et minisatellites Ces polymorphismes correspondent à l'existence d'un nombre variable de répétitions d'un motif particulier dans une séquence d'ADN. Suivant le nombre de nucléotides constituant le motif, on parlera de minisatellites (plusieurs centaines de nucléotides) ou de microsatellites (de deux à quatre nucléotides). Les microsatellites sont des variants relativement fréquents : on en trouve environ un tous les 3 à 10 kb (kilobases) (Cooper et Krawczak, 1999). Ils sont particulièrement intéressants car ils sont très polymorphes. Pour ces marqueurs, le nombre de répétitions du motif définit une forme allélique.

Les SNPs Les SNPs (*Single Nucleotide Polymorphisms*) sont des polymorphismes dus à des changements ponctuels d'un seul nucléotide. Ils sont très utilisés actuellement dans les études génétiques car ils possèdent un certain nombre de propriétés intéressantes :

- Ils sont plus abondants que les autres types de marqueurs précédemment décrits (plus de 10 millions de SNPs cartographiés sur l'ensemble du génome (Burton *et al.*, 2005)). Leur fréquence est très variable selon les régions du génome et les populations étudiées. L'abondance des SNPs permet de disposer de cartes de marqueurs très denses, ce qui est très utile pour la localisation des locus de susceptibilité ;
- Ils sont présents partout dans le génome : dans les introns, les exons, les promoteurs, les enhancers, etc. ;
- Ils sont généralement bialléliques, rarement trialléliques. Cela simplifie leur étude et, en particulier, la modélisation de leurs effets ;
- Ils sont relativement faciles à génotyper ;
- Ils mutent relativement peu. Cela signifie qu'a priori, lorsqu'un polymorphisme a muté une fois, il y a peu de chance pour qu'il subisse une réversion. On se trouve donc dans le cas du modèle de sites infinis, classique en génétique des populations. On verra par la suite que l'analyse de nos données remet en question cette affirmation.

Pour une revue concernant les SNPs, on pourra se reporter à l'article de Schork *et al.* (2000).

La notion de déséquilibre de liaison En génétique épidémiologique, beaucoup de méthodes tirent partie du déséquilibre de liaison[1] existant entre différents polymorphismes dans une population.

Définition Soient deux polymorphismes bialléliques A et B. Leurs allèles respectifs sont :

- A_1, allèle majeur de fréquence p et A_2, allèle mineur de fréquence $1-p$;
- B_1, allèle majeur de fréquence q et B_2, allèle mineur de fréquence $1-q$.

Soient $f_{A_1B_1}$, $f_{A_2B_1}$, $f_{A_1B_2}$, $f_{A_2B_2}$ les fréquences des combinaisons A_1B_1, A_2B_1, A_1B_2 et A_2B_2.

Lorsque les allèles des deux locus sont associés aléatoirement dans les gamètes, on dit que les deux locus A et B sont en *équilibre gamétique*. Les fréquences de chaque combinaison sont alors égales au produit des fréquences des allèles qui les constituent :

$$f_{A_1B_1} = pq, \quad f_{A_2B_1} = (1-p)q, \quad f_{A_1B_2} = p(1-q) \quad et \quad f_{A_2B_2} = (1-p)(1-q)$$

Si ces égalités ne sont pas respectées, on dit qu'il y a déséquilibre gamétique. Ce déséquilibre gamétique, D, se définit comme la différence entre les fréquences observées des combinaisons alléliques et les fréquences de ces combinaisons à l'équilibre.

$$\begin{array}{ll} D_{A_1B_1} = f_{A_1B_1} - pq = D & D_{A_1B_2} = f_{A_1B_2} - p(1-q) = -D \\ D_{A_2B_1} = f_{A_2B_1} - (1-p)q = -D & D_{A_2B_2} = f_{A_2B_2} - (1-p)(1-q) = D \end{array}$$

Lorsque $D_{A_1B_1}$ est positif, cela signifie que les allèles A_1 et B_1 sont plus souvent associés que ce que l'on attend si les allèles aux deux locus sont indépendants.

D'autres indices permettent de quantifier le déséquilibre gamétique :

- La mesure du D' de Lewontin et Kojima (1960) qui est moins sensible aux fréquences alléliques que D.

$$\begin{cases} D' = \frac{D}{min(pq,(1-p)(1-q))} & si\ D < 0 \\ D' = \frac{D}{min(p(1-q),(1-p)q)} & si\ D \geq 0 \end{cases}$$

  Si D' vaut 1 ou -1, cela signifie que deux allèles sont en déséquilibre complet : seules trois combinaisons alléliques existent dans la population. Si en plus les fréquences alléliques sont égales, alors seules deux combinaisons sur les quatre sont présentes dans la population ;
- Le coefficient de corrélation r^2 qui est défini par :

$$r^2 = \frac{D^2}{pq(1-p)(1-q)}$$

  Si r^2 vaut 1, les allèles sont en déséquilibre complet et seules deux combinaisons alléliques existent dans la population.

[1] Le terme de déséquilibre gamétique est plus approprié que celui de déséquilibre de liaison comme nous allons le voir dans ce paragraphe. Cependant, le terme de déséquilibre de liaison étant le plus couramment employé, c'est celui que nous retiendrons pour la suite de ce travail

Genèse et évolution d'un déséquilibre gamétique Un déséquilibre gamétique entre les allèles aux locus A et B peut être créé par différents mécanismes :

- Lorsqu'une mutation se produit dans la lignée germinale d'un individu, le nouvel allèle créé va se trouver systématiquement associé aux allèles des autres polymorphismes ;
- Lorsqu'une combinaison d'allèles est sélectionnée, sa fréquence va augmenter dans la population, générant un déséquilibre ;
- Un mélange de populations ayant des fréquences alléliques différentes aux deux locus peut aussi générer un déséquilibre, par exemple, dans le cas extrême du mélange de deux populations homogènes, l'une constituée uniquement d'individus $A_1A_1B_1B_1$ et l'autre uniquement d'individus $A_2A_2B_2B_2$. Dans la population résultante, on peut trouver uniquement deux types de gamètes (A_1B_1 et A_2B_2) parmi les quatre possibles.

On voit ici que le déséquilibre gamétique n'existe pas forcément uniquement entre des locus liés génétiquement (i.e. portés par le même chromosome). C'est pour cette raison que le terme déséquilibre de liaison, couramment employé pour désigner le déséquilibre gamétique est ambigu.

L'évolution du déséquilibre entre deux marqueurs au cours du temps va cependant dépendre de leur liaison génétique. En effet, il est possible de montrer que, dans une population, le déséquilibre va décroître d'autant plus vite que les deux locus sont éloignés. Si l'on suppose un déséquilibre initial D_0 entre deux locus et un taux de recombinaison θ, en l'absence de création de nouveau déséquilibre au cours des générations suivantes, le déséquilibre à la génération n vaudra :

$$D_n = D_0(1-\theta)^n$$

Dans le cas de locus non liés ($\theta = 0,5$), le déséquilibre initial diminue de moitié à chaque génération alors que, si les locus sont très proches ($\theta \simeq 0$), le déséquilibre se maintient. On s'attend donc à ne trouver du déséquilibre qu'entre des locus relativement proches. Cependant, du fait des nombreux mécanismes générateurs de déséquilibre qui peuvent potentiellement se produire au cours de l'évolution, il a été observé que, dans certaines populations, cette relation avec la distance n'est pas stricte (Bourgain *et al.* (2000)).

Utilisation conjointe de plusieurs polymorphismes : notion d'haplotypes Un haplotype correspond à l'ensemble des allèles de différents marqueurs situés sur le même brin d'ADN. Pour n SNPs bialléliques, il y a donc, en théorie, 2^n haplotypes possibles. En pratique, du fait du déséquilibre de liaison, on en observe toujours beaucoup moins.

L'un des problèmes majeurs concernant les haplotypes est qu'ils ne sont pas directement observables pour des individus diploïdes. Lors du séquençage, on a seulement accès aux génotypes non phasés, c'est-à-dire qu'on connaît les deux allèles aux différents locus mais qu'on ne sait pas quels allèles sont situés sur le même chromosome. Pour les individus homozygotes à tous les locus ou hétérozygotes à un seul locus, il est possible d'obtenir les paires d'haplotypes des individus à partir de leurs génotypes non phasés. Cependant, dès que l'individu est hétérozygote à plus de deux locus, la reconstruction est ambiguë car plusieurs paires d'haplotypes sont compatibles avec ce génotype. Ces ambiguïtés peuvent parfois être résolues grâce à l'apport d'information familiale.

Il existe plusieurs techniques de génotypage moléculaire (pour une revue, voir Niu (2004)) mais ces méthodes sont très lourdes à mettre en place et extrêmement coûteuses.

Elles ne sont donc pas adaptées aux analyses actuelles qui prennent en compte de plus en plus de marqueurs.

Si l'on ne dispose pas de données sur les parents ou si l'information familiale est insuffisante, il faut alors envisager tous les haplotypes possibles pour un individu et estimer leurs probabilités.

De nombreux algorithmes ont été développés pour inférer les haplotypes à partir des génotypes non phasés.

Reconstruction haplotypique sur des données en population Beaucoup de programmes de reconstruction haplotypique sont basés sur des algorithmes d'espérance-maximisation (EM) (Dempster *et al.*, 1977). Des valeurs arbitraires de fréquence sont tout d'abord affectées aux différents haplotypes possibles et sont utilisées pour estimer les fréquences génotypiques (étape E : Espérance). Les valeurs des fréquences génotypiques sont à leur tour utilisées pour estimer les fréquences haplotypiques à l'itération suivante (étape M : Maximisation) et ainsi de suite, jusqu'à ce qu'il y ait convergence[2]. Cependant, ces algorithmes supposent que les hypothèses de Hardy-Weinberg sont vérifiées pour l'ensemble des haplotypes, ce qui n'est pas forcément le cas, par exemple, pour un échantillon de malades.

D'autres algorithmes sont basés sur l'utilisation de méthodes bayésiennes, ou pseudo-bayésiennes comme, par exemple, l'algorithme implémenté dans le logiciel PHASE (Stephens *et al.*, 2001a; Stephens et Donnelly, 2003). Les méthodes bayésiennes utilisent à la fois une information *a priori* concernant les fréquences haplotypiques et l'information contenue dans le jeu de données (i.e. les génotypes), pour calculer la distribution *a posteriori* des fréquences haplotypiques sachant les génotypes observés dans l'échantillon. Dans le cas de PHASE , l'algorithme utilise un processus de MCMC (*Monte Carlo Markov Chain*) et la distribution *a priori* est obtenue en utilisant la théorie de la coalescence. En pratique, ce dernier point signifie que lorsque plusieurs paires d'haplotypes sont compatibles avec un génotype donné, on donne un poids plus important aux haplotypes déjà présents (ou supposés être présents) dans l'échantillon. Si un haplotype n'existe pas dans l'échantillon, alors on lui attribue un poids d'autant plus important qu'il est proche (en terme de nombre de pas évolutifs) d'un autre haplotype de l'échantillon.

Enfin, il existe des algorithmes basés sur la reconstruction phylogénétique : les haplotypes sont reconstruits de telle manière qu'un arbre phylogénétique parfait[3] (Gusfield, 2002) ou imparfait (Halperin et Eskin, 2004) raconte leur histoire évolutive. Pour une revue sur les méthodes de reconstructions d'haplotypes, voir Niu (2004).

Les performances respectives des méthodes basées sur des algorithmes EM et des différentes méthodes bayésiennes ont été étudiées par différents auteurs. Adkins (2004) a montré que toutes ces méthodes sont globalement très efficaces pour estimer les fréquences haplotypiques. Malgré quelques contestations (Zhang *et al.*, 2001), il semble cependant que PHASE donne les reconstructions haplotypiques les plus précises (Xu *et al.*, 2004; Stephens et Donnelly, 2003; Stephens *et al.*, 2001b). Bien que le temps de calcul de PHASE soit beaucoup plus long que celui des autres algorithmes, il a été choisi pour reconstruire les haplotypes

[2]On dit qu'il y a convergence lorsque les différences entre les fréquences haplotypiques obtenues pour deux itérations successives sont très faibles

[3]Une phylogénie est parfaite lorsqu'il n'y a pas d'homoplasie. Cela signifie qu'il ne doit pas y avoir de recombinaison et que le modèle site-infini doit être respecté

dans les bases de données du *Variation discovery resource project* (SeattleSNPs) et du projet HapMap(Consortium, 2003).

Reconstruction haplotypique sur des données familiales Les données familiales apportent un supplément d'information : seules certaines paires d'haplotypes sont compatibles avec le génotype d'un individu connaissant celui de ses parents. Malgré cela, peu de méthodes de reconstructions haplotypiques ont été développées pour les données familiales. Les programmes d'analyse de liaison tels que GENEHUNTER ont été souvent utilisés pour inférer les haplotypes dans des familles. Cependant, Schaid *et al.* (2002) ont montré que l'utilisation de ces programmes n'était pas appropriée car ils supposent qu'il n'y a pas de déséquilibre de liaison entre les marqueurs. Aussi, lorsque la reconstruction haplotypique est ambiguë, les différents haplotypes possibles sont tous aussi vraisemblables les uns que les autres. Le programme attribuera alors à l'individu une paire d'haplotypes en fonction de la numérotation des marqueurs. Quelques autres algorithmes ont donc été développés mais assez peu ont été implémentés sous la forme de logiciels. Il existe actuellement essentiellement l'algorithme implémenté dans le logiciel FAMHAP Becker et Knapp (2004a). Ce programme est basé sur un algorithme EM, adapté pour les données familiales [4]. Grâce à une procédure d'addition pas à pas des locus, il permet de reconstruire des haplotypes pour un nombre élevé de marqueurs (jusqu'à 68) en un temps relativement court.

1.1.1.2 La notion de pénétrance

La pénétrance est l'un des paramètres qui permet de spécifier la relation entre un phénotype (malade/non malade) et le génotype à un locus particulier. Soit un locus de susceptibilité biallélique dont l'allèle à risque est noté d. La pénétrance est définie comme la probabilité pour un individu d'être atteint de la maladie sachant son génotype au locus de susceptibilité. Dans le cas d'un locus biallélique, on définit alors trois pénétrances :

$$\begin{cases} f_{DD} = P(malade|DD) \\ f_{Dd} = P(malade|Dd) \\ f_{dd} = P(malade|dd) \end{cases}$$

Lorsqu'un facteur de risque génétique a un effet dominant, alors $f_{Dd} = f_{dd}$ et lorsque le facteur a un effet récessif, $f_{Dd} = f_{DD}$.

Pour certaines maladies, la correspondance entre le phénotype et le génotype est simple. Par exemple, pour une maladie monogénique dominante, tous les individus portant au moins un allèle d sont malades ($f_{Dd} = f_{dd} = 1$ et $f_{DD} = 0$). Dans les autres cas, on parlera de :

- maladie à *pénétrance incomplète* si $f_{dd} < 1$. Cela signifie qu'il existe des individus non malades homozygotes pour l'allèle à risque
- *phénocopies* si $f_{DD} > 0$. Cela signifie qu'il existe des individus malades non porteurs de l'allèle à risque, la maladie étant due à d'autres facteurs, soit génétiques, soit environnementaux.

Dans le cas des maladies multifactorielles, les pénétrances sont rarement complètes et on observe très souvent des phénocopies. Cette difficulté d'établir une correspondance stricte entre le génotype et le phénotype est l'une des difficultés de l'étude de ce type de maladies.

[4] Le logiciel FAMHAP permet cependant aussi de reconstruire les haplotypes de données en population

Au lieu de spécifier les trois valeurs de pénétrance, on utilise souvent la notion de risque relatif génotypique. Cela consiste à exprimer les risques de développer la maladie pour les individus homozygotes porteurs de l'allèle à risque et pour les individus hétérozygotes par rapport au risque des individus homozygotes non porteurs.

$$GRR_{hom} = \frac{\text{pénétrance des homozygotes porteurs}}{\text{pénétrance des homozygotes non porteurs}} = \frac{f_{dd}}{f_{DD}}$$

$$GRR_{het} = \frac{\text{pénétrance des hétérozygotes}}{\text{pénétrance des homozygotes non porteurs}} = \frac{f_{dD}}{f_{DD}}$$

1.1.1.3 Les méthodes de recherche de facteurs de risque génétiques

Il existe deux grandes stratégies d'identification de facteurs de risque génétiques : les analyses de liaison et les analyses d'association. Pour une revue de ces grands types d'analyse, voir Lander et Schork (1994). Un troisième groupe de méthodes, les méthodes de recherche d'haplotypes partagés, est parfois défini car il emprunte des concepts à la fois aux analyses d'association et aux analyses de liaison.

Les analyses de liaison Les analyses de liaison ont été développées pour localiser des facteurs génétiques de risque sur des régions chromosomiques précises. Elles se sont révélées particulièrement efficaces dans le cas de l'étude des maladies monogéniques. Elles consistent à rechercher des marqueurs dont la transmission dans des familles n'est pas indépendante de la transmission de la maladie. Ces marqueurs seront donc situés dans des régions proches de celle du locus de susceptibilité. En effet, lorsque deux locus sont situés dans une même région de chromosome, on s'attend à ce qu'il se produise peu de recombinaisons entre eux et donc qu'ils soient transmis ensemble dans des familles plus souvent que s'ils n'étaient pas liés. Les méthodes d'analyse de liaison sont de deux types :

- Les méthodes paramétriques : elles nécessitent la spécification d'un modèle génétique pour la maladie. Elles sont basées sur la statistique du *lod score*. Elles sont relativement puissantes tant que le modèle génétique spécifié est correct. Lorsque le modèle est mal spécifié, Clerget-Darpoux *et al.* (1986) ont montré qu'une région contenant un locus de susceptibilité pouvait être exclue à tort. Dans le cas des maladies multifactorielles, le modèle n'étant pas connu, le risque de fausse exclusion est relativement élevé. Il se double d'un risque de fausse détection car, en l'absence de modèle connu, les utilisateurs sont tentés de tester un grand nombre de modèles, ce qui augmente la probabilité d'observer un résultat significatif par hasard. Pour plus de détails sur ces méthodes, voir Feingold *et al.* (1998).
- Les méthodes "modèles-indépendantes" : elles ne nécessitent pas la spécification d'un modèle génétique pour la maladie. Elles consistent à rechercher des régions chromosomiques identiques chez des individus apparentés atteints par une maladie (souvent, il s'agit de paires de germains atteints). En effet, si un gène est impliqué dans la susceptibilité à une maladie, on s'attend à ce que deux individus apparentés malades aient plus souvent reçu les mêmes allèles aux marqueurs liés au locus maladie que ne le voudrait le hasard de la ségrégation. Ces méthodes sont plus robustes que les méthodes paramétriques, mais elles sont aussi moins puissantes.

Les méthodes de recherche d'haplotypes partagés Le principe de ces méthodes consiste à rechercher des régions sur le chromosome où les individus malades sont plus apparentés que les témoins, c'est-à-dire qu'ils partagent plus d'allèles identiques hérités d'un ancêtre commun. Ces régions sont celles qui ont le plus de risque d'être impliquées dans la susceptibilité à la maladie. Une mesure de l'apparentement fréquemment utilisée dans ces méthodes est la longueur d'haplotype partagée. En effet, plus la longueur partagée est grande, plus il y a de chance pour que les individus qui partagent cette région l'ait héritée d'un ancêtre commun. Au contraire, si la région est courte, il y a de plus grandes chances que l'identité soit due au hasard.

Il existe plusieurs méthodes de ce type. Nous pouvons citer par exemple la méthode HSS (Van der Meulen et te Meerman, 1997) et la méthode MILC (Bourgain *et al.*, 2000, 2001). Cependant, Bourgain *et al.* (2002) ont montré que pour de nombreux modèles génétiques, ces méthodes ne sont pas plus puissantes que des tests d'association classique.

Les analyses d'association Les tests d'association cherchent à identifier des régions impliquées dans le déterminisme d'une maladie. Pour cela, ils utilisent l'information apportée soit directement par le locus de susceptibilité s'il correspond à l'un des marqueurs étudiés, soit par des marqueurs en déséquilibre de liaison avec le locus de susceptibilité. Contrairement aux analyses de liaison qui utilisent l'information des transmissions familiales, les analyses d'association sont basées sur la comparaison d'échantillons de cas (individus atteints par une maladie) et de témoins (individus non malades).

Le principe général de ces tests consiste à comparer les distributions alléliques ou génotypiques de différents marqueurs chez les malades et les témoins. Si ces distributions diffèrent, cela signifie que le marqueur considéré est associé à la maladie. Ce marqueur peut donc être soit lui-même le site de susceptibilité à la maladie, ou, plus généralement, être en déséquilibre de liaison avec le site de susceptibilité.

Les analyses d'association peuvent se réaliser sur des polymorphismes répartis sur l'ensemble du génome (approche systématique) comme c'est le cas pour les analyses de liaison. Cependant, étant donné que le déséquilibre de liaison n'existe qu'entre des marqueurs très proches, il faudra utiliser un très grand nombre de marqueurs dans ces analyses (plusieurs centaines de milliers, Risch (2000)). Dans ce cas, le seuil de significativité du test d'association devra être très élevé afin d'éviter de détecter des faux positifs. Par conséquent, il sera difficile de mettre en évidence des facteurs ayant un effet modéré.

Une solution à ce problème est de ne pas utiliser tous les marqueurs disponibles, mais seulement ceux qui sont les plus informatifs. Par exemple, il est évident que si deux marqueurs sont en déséquilibre de liaison complet ($r^2 = 1$), il est inutile de les analyser tous les deux car ils apportent la même information. Dans certaines régions du génome, on a pu mettre en évidence des blocs de marqueurs en très fort déséquilibre de liaison (Daly *et al.*, 2001b; Gabriel *et al.*, 2002; Patil *et al.*, 2001). On peut alors choisir un SNP comme représentant de chaque bloc et ne réaliser le test d'association que sur ces SNPs, appelés alors ht-SNP (*haplotype tagging SNPs*). Cependant, il faut noter que ce découpage en bloc n'est pas absolu, en particulier parce que les frontières entre les blocs sont difficiles à définir. En effet, il existe différentes méthodes de détection des blocs, basées sur le déséquilibre de liaison entre chaque paire de SNPs (Gabriel *et al.*, 2002) ou sur la maximisation de la diversité haplotypique (Patil *et al.*, 2001). Schwartz *et al.* (2003) et Schulze *et al.* (2004) ont montré que les limites des blocs ne sont pas concordantes lorsque différentes méthodes sont utilisées.

De plus, Schulze *et al.* (2004) ont montré que les limites de bloc sont très sensibles aux paramètres de la méthode utilisée (seuil de déséquilibre de liaison par exemple) ainsi qu'à la fréquence minimale de l'allèle mineur utilisée pour sélectionner les SNPs qui vont servir pour établir les limites de bloc. Enfin, Wall et Pritchard (2003) ont montré que la structure en bloc n'était pas généralisable à l'ensemble du génome. Une telle approche n'est donc pas réalisable tant qu'une carte précise des blocs n'a pas été établie pour la population étudiée et sa puissance est encore controversée (Zhai *et al.*, 2004). Enfin, même si une association est définie entre un bloc et une maladie, d'autres outils seront nécessaires afin d'identifier plus précisément les polymorphismes à risque dans le bloc.

En plus des méthodes basées sur le déséquilibre de liaison ou sur la diversité haplotypique, il existe de nombreuses autres méthodes de sélection de SNPs dont, par exemple, :

- La sélection sur la fréquence de l'allèle mineur. On ne conserve que les SNPs dont la fréquence de l'allèle mineur est supérieure à un seuil (souvent 5 ou 10%) ;
- La sélection sur la région contenant les marqueurs. On peut choisir de ne s'intéresser qu'aux SNPs dont les mutations ont des chances d'avoir un effet sur le gène ou sur son niveau d'expression (dans les exons, dans les promoteurs, etc.),

Cependant, ces méthodes sont aussi très dépendantes de leurs différents paramètres (choix des régions, fréquences des allèles mineurs) et l'impact des variations de ces paramètres sur leur puissance n'a jamais été quantifié.

Une stratégie alternative est de se focaliser sur des gènes particuliers — dits gènes *candidats* — et de tester s'ils sont impliqués dans la susceptibilité à la maladie. Le choix des gènes peut être basé sur leur fonction, sur des homologies avec des gènes qui induisent des signes proches de ceux de la maladie lorsqu'ils sont modifiés dans des modèles animaux ou simplement sur le fait que le gène se trouve dans une région dont on a montré la liaison avec la maladie. Dans la suite de ce travail, nous allons nous intéresser uniquement à la stratégie de gène candidat car elle semble bien adaptée à l'étude des maladies multifactorielles.

1.1.2 Les méthodes d'association en stratégie gène candidat

1.1.2.1 Les différents tests d'association

Il existe une très grande variété de tests d'association utilisés en stratégie gène candidat. Nous allons à présent en détailler les principaux pour le cas d'analyses de données qualitatives (données cas/témoins).

Les tests univariés Les tests les plus simples et les plus couramment utilisés consistent à tester l'homogénéité des distributions d'allèles ou de génotypes chez les malades et les témoins. Ces tests étant réalisés successivement pour tous les marqueurs du gène, on les appellera par la suite tests "locus par locus". Si l'on considère un marqueur biallélique (A_1, A_2), les données peuvent être résumées sous la forme d'un tableau de contingence comme illustré sur le Tableau 1.1).

L'homogénéité entre les distributions génotypiques chez les malades et chez les témoins est alors testée en calculant un χ^2 dont le nombre de degrés de liberté (ddl) sera égal au nombre de génotypes moins 1 (ici, 2 ddl). Ce test est réalisé à chaque SNP étudié. Il est donc nécessaire de tenir compte des tests multiples. Ce type de test est particulièrement efficace lorsqu'un seul variant est responsable de la susceptibilité à la maladie. Dans ce cas,

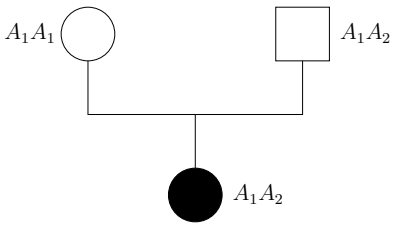
	A_1A_1	A_1A_2	A_2A_2	Total
Malades	$m_{1,1}$	$m_{1,2}$	$m_{2,2}$	M
Témoins	$t_{1,1}$	$t_{1,2}$	$t_{2,2}$	T
Total	$N_{1,1}$	$N_{1,2}$	$N_{2,2}$	N

TAB. 1.1: Exemple de table de contingence pour un test d'association génotypique

la puissance du test va dépendre du déséquilibre de liaison qui existe entre le marqueur étudié et le site de susceptibilité, de l'effectif total et des pénétrances au site de susceptibilité.

Comme toutes les études cas/témoins, les analyses d'association sont confrontées au problème du choix des témoins. En particulier, il est essentiel que les cas et les témoins proviennent de la même population. En effet, nous avons vu précédemment qu'un mélange de populations pouvait créer un déséquilibre de liaison si les fréquences des allèles sont différentes dans les deux populations. Or, les études d'association testent l'existence d'un déséquilibre de liaison et l'interprètent comme l'indication que le marqueur étudié est proche du locus de susceptibilité. Si le déséquilibre est créé par l'échantillonnage des malades et des témoins, cette interprétation est alors fausse et on conclura à tort à l'existence d'une association entre le marqueur et la maladie. Pour pallier ce problème, Falk et Rubinstein (1987) ont proposé d'utiliser des témoins internes (méthode HRR pour *Haplotype Relative Risk*) : on va s'intéresser aux parents des individus malades et on va utiliser comme témoins les haplotypes ou les allèles des parents qui ne sont pas transmis aux enfants malades.

Plus récemment, Spielman *et al.* (1993) ont proposé un test permettant de tester à la fois la liaison et l'association entre le marqueur et la maladie : le TDT (*Transmission Disequilibrium Test*). Pour un trait qualitatif, le test consiste à comparer la distribution des allèles transmis et non transmis par des parents hétérozygotes à un enfant atteint à la distribution attendue sous l'hypothèse nulle (même probabilité de transmission quel que soit l'allèle). En cas de liaison entre le marqueur et la maladie, on observera que l'allèle à risque sera plus souvent transmis par les parents aux enfants atteints. Le test est détaillé sur la Figure 1.1.



Dans la famille représentée ci-contre, seul le père sera informatif pour le TDT. Il transmet à son enfant malade l'allèle A_2.
Soient N parents hétérozygotes dont x ont transmis l'allèle A_1 et y l'allèle A_2. Sous l'hypothèse nulle, on s'attend à ce que $\frac{N}{2}$ allèles A_1 et $\frac{N}{2}$ allèles A_2 soient transmis. Le TDT consiste à comparer x et y à $\frac{N}{2}$ par un test d'homogénéité.

FIG. 1.1: Le *Transmission Disequilibrium Test*

Il existe de nombreuses extensions du TDT qui permettent de l'utiliser dans pratiquement toutes les situations (données quantitatives, utilisation de marqueurs multi-alléliques, etc.).

Les analyses multivariées Les analyses multivariées permettent de prendre en compte l'information de plusieurs SNPs. Il existe de nombreuses méthodes d'analyse multivariée. Nous ne détaillerons ici que celles qui seront évoquées dans la suite de notre travail.

Les méthodes de régression Il est possible de modéliser la relation entre la maladie et les différents marqueurs par un modèle de régression. Selon le type de données, il s'agira de régression linéaire (données quantitatives) ou logistique (données qualitatives). Dans le cas de la régression logistique, cela consiste à écrire que la probabilité P_i pour un individu i d'être atteint dépend de l'effet individuel des génotypes à chaque marqueur et, éventuellement, d'interactions entre les marqueurs. Dans le cas de SNPs bialléliques (0/1), si l'on ne considère que les termes d'interaction du second degré (i.e. entre 2 SNPs, notés $\beta_{k,l}$), on a :

$$logit(P_i) = log(\frac{P_i}{1-P_i}) = \beta_0 + \sum_{k=1}^{n} \beta_k X_{i,k} + \sum_{1 \leq k < l \leq n} \beta_{k,l} X_{i,k} X_{i,l}$$

où n est le nombre de marqueurs, et où les $X_{i,k}$ représentent le statut de l'individu i pour le SNP k ($X_{i,k}$=0, 1,-1 respectivement pour les génotypes 11, 12 et 22). Pour simplifier le problème, on peut se placer dans un cas où il n'y a pas d'interaction ($\beta_{k,l} = 0$). Dans ce cas, les β_k correspondent au logarithme du risque relatif pour les différents génotypes. On pourra alors tester si certains des β sont nuls, ce qui indique que ces SNPs n'ont pas d'effet sur le statut maladie.

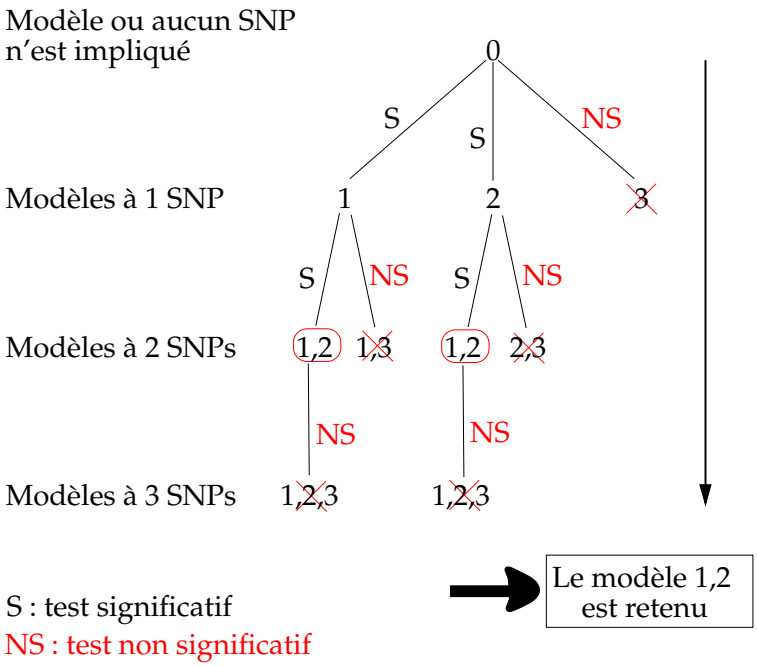
Pour rechercher les SNPs qui sont impliqués dans le déterminisme d'une maladie, on peut utiliser une procédure pas à pas qui compare la vraisemblance de modèles incluant k SNPs à celle de modèles incluant $k+1$ SNPs —k variant de 1 à $n-1$ — afin d'identifier celui qui explique le mieux les données (Cordell et Clayton, 2002). La procédure de test peut être ascendante (on ajoute petit à petit des locus au modèle) ou descendante (on enlève petit à petit des locus au modèle), voire, combiner les deux. Le test de comparaison de modèle est un test de rapport de vraisemblance : sous l'hypothèse nulle, le double de la différence entre les vraisemblances des deux modèles suit approximativement une loi de χ^2 dont le nombre de degrés de liberté correspond à la différence entre le nombre de paramètres estimés pour chaque modèle. La Figure 1.2 détaille le processus pour une analyse ascendante.

Nous avons expliqué ici le principe de la régression logistique appliquée à des données de type cas/témoins. Cependant, il est tout à fait possible d'étendre cette méthode à des analyses de données familiales. Pour cela, lors du calcul des vraisemblances, il faut prendre en compte le fait que l'observation du génotype des enfants est conditionnée par le statut maladie *et* par les génotypes des parents (Cordell et Clayton, 2002).

La méthode de régression pas à pas présente l'avantage de ne pas réaliser tous les tests possibles, ce qui permet de ne pas trop diminuer la puissance de la méthode. De plus, les tests réalisés ont souvent un nombre assez faible de degrés de liberté, ce qui permet aussi de conserver plus de puissance qu'un test classique. Cependant, cette approche présente l'inconvénient de ne pas intégrer de correction pour les tests multiples. En effet, comme les tests réalisés à un niveau dépendent des résultats des tests au niveau précédent, il n'est pas possible de prévoir *a priori* combien de tests et quels tests seront réalisés. Il faut donc être très prudent lors de l'interprétation des résultats, surtout lorsque les $p-values$ sont proches du seuil de signification choisi (5% ou 1%).

Le test de combinaison Ce test a été développé indépendamment par Jannot *et al.* (2003) et par Becker et Knapp (2004b) qui l'ont implémenté dans le programme FAMHAP . Il consiste à tester l'association entre les génotypes (ou les haplotypes, ou les diplotypes[5]) et

[5] un diplotype correspond à l'ensemble des deux haplotypes d'un individu



Au premier niveau de l'analyse, on teste pour chaque SNP si un modèle avec un effet de ce SNP est plus vraisemblable qu'un modèle où aucun SNP n'a d'effet. Sur la figure, c'est le cas des modèles comprenant respectivement les SNPs 1 et 2 ;

Pour tous les modèles à un locus qui ont donné un test significatif, on teste si l'addition d'un second locus donne un modèle plus vraisemblable. Sur la figure, c'est le cas du modèle impliquant les SNPs 1 et 2 ensemble ;

De même, sur les modèles à 2 locus retenus au niveau 2, on teste si l'addition d'un troisième locus donne un modèle plus vraisemblable. Ici, ce n'est pas le cas. Le modèle qui est conservé est le modèle à deux locus 1,2.

FIG. 1.2: Principe de la régression logistique ascendante pas à pas

la maladie pour toutes les combinaisons possibles formées par un nombre variable de SNPs. Par exemple, dans le cas de l'étude de 4 SNPs, on va réaliser successivement :

- un test pour chaque SNP → 4 tests
- un test pour chaque combinaison de 2 SNPs → 6 tests
- un test pour chaque combinaison de 3 SNPs → 4 tests
- un test pour l'ensemble des 4 SNPs

Au final, 15 tests sont réalisés. Dans le cas d'une analyse cas/témoins, il s'agit de tests de χ^2 dont le nombre de degrés de liberté correspond au nombre de génotypes (ou d'haplotypes ou de diplotypes) moins un. Dans le cas de données familiales (trio), il s'agira de tests de TDT (génotypique ou haplotypique).

On remarque que plus le nombre de SNPs étudié est grand, plus le nombre de tests est important. Il va donc être nécessaire de tenir compte des tests multiples pour limiter les risques de détecter des faux positifs. Cependant, les tests étant très dépendants, il faudra veiller à ne pas trop corriger pour ne pas perdre trop de puissance. Pour cela, la procédure de permutation de Ge *et al.* (2003) détaillée page 20 est utilisée. Enfin, il faut noter que si le

nombre de SNPs dépasse 15 ou 20, le temps de calcul devient très long.

Les tests haplotypiques

Les tests de distribution haplotypique Ils testent l'homogénéité de la distribution des haplotypes chez les malades et les témoins. Soit un ensemble de k catégories d'haplotypes[6]. La méthode la plus simple consiste à tester l'homogénéité en utilisant un test de χ^2 dont le nombre de degrés de liberté sera égal à $k-1$. Dans le cas de données familiales, une variante du TDT, nommée TDT_{all} peut être utilisée. Le principe est alors de tester la symétrie de la Table 1.2.

Haplotypes	non transmis							Total
transmis	H_1	H_2	...	H_i	...	H_k		
H_1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,i}$	...	$n_{1,k}$		$n_{1,.}$
H_2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,i}$	...	$n_{2,k}$		$n_{2,.}$
...	...	...	...	...		...	...	...
H_j	$n_{j,1}$	$n_{j,2}$	...	$n_{j,i}$	...	$n_{j,k}$		$n_{j,.}$
...	...	...	...	...		.....	...	
H_k	$n_{k,1}$	$n_{k,2}$	...	$n_{k,i}$	...	$n_{k,k}$		$n_{k,.}$
Total	$n_{.,1}$	$n_{.,2}$	...	$n_{.,i}$	...	$n_{.,k}$		$n_{.,.}$

Les $n_{i,j}$ correspondent aux nombres de parents hétérozygotes H_iH_j qui ont transmis l'haplotype H_i à un enfant atteint (l'haplotype H_j n'a donc pas été transmis à cet enfant).

TAB. 1.2: Principe du TDT_{all}

Test de l'effet d'un haplotype par rapport aux autres Au lieu de tester d'un bloc l'ensemble des haplotypes, l'effet de chaque catégorie d'haplotypes peut être testé tour à tour contre l'effet de l'ensemble des autres catégories haplotypiques, ce qui permet d'identifier les haplotypes associés à la maladie. Si k catégories d'haplotypes sont considérées, k tests à un degré de liberté seront réalisés au lieu d'un test à $k-1$ degrés de liberté. Dans le cas de données familiales, une méthode de ce genre, appelée TDT_{max} a été mise au point par Clayton (1999) et implémentée dans le programme `TRANSMIT`.

Cette méthode est particulièrement efficace lorsqu'un seul haplotype est impliqué dans la susceptibilité à la maladie. Quand ce n'est pas le cas, des catégories d'haplotypes dans lesquelles les distributions de malades et de témoins ne sont pas homogènes seront considérées ensemble, ce qui fait baisser la puissance du test. De plus, comme plusieurs tests sont réalisés, il est nécessaire de tenir compte des tests multiples.

[6]Le terme "haplotype" est ambigu : il peut soit désigner une entité unique (un individu porte deux haplotypes), soit représenter un ensemble d'haplotypes identiques dans un échantillon. Dans ce dernier cas, on parlera généralement de catégorie haplotypique afin de limiter les confusions

Intérêt relatif des tests locus par locus et des tests haplotypiques L'avantage des méthodes haplotypiques par rapport aux tests individuels est un sujet très débattu dans la littérature. Certaines études montrent que les tests haplotypiques sont plus puissants (Akey *et al.*, 2001; Zaykin *et al.*, 2002; Zhang *et al.*, 2002), d'autres concluent que les analyses individuelles sont préférables (Long et Langley, 1999; Kaplan et Morris, 2001; Roeder *et al.*, 2005). Finalement, il semble que la puissance relative de ces deux approches dépende de plusieurs facteurs :

- la présence ou non du SNP à risque parmi l'ensemble des marqueurs étudiés (Judson et Stephens, 2001; Bader, 2001) ;
- Le nombre de sites de susceptibilité (Bader, 2001) ;
- Le nombre d'allèles à risque au locus de susceptibilité (Morris et Kaplan, 2002) dans le cas de locus multi-alléliques
- Le modèle de la maladie et les éventuelles interactions entre les sites de susceptibilité (Culverhouse *et al.*, 2002; Jannot *et al.*, 2003).

1.1.2.2 Les méthodes de localisation des sites de susceptibilité

Lorsqu'une association a été détectée entre une région du génome et une maladie, on cherche alors à localiser de façon précise les locus de susceptibilité à la maladie. Selon les méthodes, on va soit identifier directement un ou plusieurs marqueurs à risque, soit localiser une région, la plus petite possible, contenant le locus de susceptibilité.

Identification directe du site de susceptibilité Les tests locus par locus et les analyses multivariées permettent de faire des hypothèses concernant le ou les locus de susceptibilité. Dans le cas des tests locus par locus, le ou les locus les plus associés à la maladie sont des sites de susceptibilité potentiels. Les méthodes multivariées permettent d'identifier le SNP ou la combinaison de SNPs qui est la plus associée à la maladie (cas du test de combinaison), ou qui explique le mieux les statuts malades ou non malades des individus (cas des méthodes de).

Identification d'une région sur le chromosome contenant le site de susceptibilité Ces méthodes sont héritées des méthodes de recherche de gènes impliqués dans les maladies monogéniques. Dans ce cas, la maladie est due à une mutation unique qui a eu lieu sur un haplotype donné. En fonction de l'âge de la mutation, les malades vont donc partager des segments plus ou moins grands de l'haplotype portant la mutation.

Dans le cas des maladies multifactorielles, l'hypothèse d'une mutation unique n'est pas correcte, mais on peut tout de même supposer que les malades possèdent un plus petit nombre d'haplotypes ancêtres que les témoins. Ainsi, on s'attend à ce que, aux environs du locus de susceptibilité, ils partagent plus d'allèles que les haplotypes des témoins. Du fait des recombinaisons que subissent les haplotypes, cette zone de partage d'allèles est d'autant plus courte que le temps d'évolution depuis l'ancêtre est long. Afin d'augmenter leur efficacité, ces méthodes sont souvent mises en œuvre dans des populations à effet fondateur, c'est-à-dire, des populations fondées par un petit nombre d'individus et restées isolées pendant la phase de croissance suivant la fondation. La population étant fondée par un nombre restreint d'individus, il est donc probable que les haplotypes des malades descendent d'un petit nombre d'ancêtres (voire d'un seul ancêtre idéalement).

La plupart des méthodes de localisation fine sont basées sur la modélisation de l'évolution des haplotypes des malades au cours des générations pour toutes les positions possibles du locus de susceptibilité en utilisant un processus de coalescence. Elles diffèrent essentiellement par la façon dont l'évolution des haplotypes est modélisée et, en particulier par la structure de la généalogie des haplotypes des malades. Selon les auteurs, elle sera en étoile[7] (McPeek et Strahs, 1999; Morris *et al.*, 2000; Liu *et al.*, 2001), avec uniquement des bifurcations (Rannala et Reeve, 2001) ou en plusieurs morceaux[8] (Morris *et al.*, 2002). La méthode de Zöllner et Pritchard (2005) se distingue des autres en ce qu'elle modélise aussi l'évolution des haplotypes des témoins. Ces méthodes présentent un inconvénient : elles nécessitent d'estimer un très grand nombre de paramètres. De ce fait, leur temps de calcul est très long.

Plus récemment, des méthodes bayésiennes ont été développées (Molitor *et al.*, 2003a,b), pour une revue voir Thomas *et al.* (2003). Ces méthodes regroupent les haplotypes sur la base de leur similarité deux à deux (Molitor *et al.*, 2003a) ou de leur distance par rapport à un "centre" qui peut être un haplotype de l'échantillon ou non (Molitor *et al.*, 2003b). En choisissant une mesure de distance (ou de similarité) faisant intervenir la position du locus de susceptibilité, cette dernière est alors un paramètre du modèle. On pourra donc obtenir une distribution de la probabilité de présence d'un locus de susceptibilité sur la région étudiée.

Les efficacités relatives de ces différentes méthodes n'ont été comparées que par Morris *et al.* (2002). Dans leur étude, ils n'ont considéré que les approches modélisant l'évolution des haplotypes des malades. Ils ont réalisé une étude par simulations où ils ont fait varier différents paramètres : le taux de recombinaison, la fréquence de l'allèle à risque et le taux de phénocopies. Ils ont montré que, dans tous les cas, l'utilisation de généalogies en étoile est moins puissante, surtout lorsque le taux de recombinaison est élevé et que l'utilisation de généalogies en plusieurs morceaux donne les meilleurs résultats.

La plupart des auteurs de ces études ont validé leur méthode en utilisant le jeu de données de Kerem *et al.* (1989) concernant la mucoviscidose et le gène CFTR (*Cystic Fibrosis Transmembrane conductance Regulator*). La position du locus à risque étant connue sur ce gène, il est donc possible de comparer leur efficacité pour localiser le site de susceptibilité sur cet exemple. Les résultats obtenus sont résumés dans le Tableau 1.3.

Ces résultats confirment ceux des simulations de Morris *et al.* (2002) : leur méthode localise mieux le locus de susceptibilité que les méthodes développées précédemment. La méthode de Liu *et al.* (2001) donne toutefois le plus faible intervalle de probabilité. La méthode de Zöllner et Pritchard (2005) est, sur ce jeu de données, la méthode la plus efficace parmi toutes celles basées sur des modèles de coalescence. Les méthodes de regroupement basées sur des distances (Molitor *et al.*, 2003a,b) donnent aussi de très bons résultats. Cependant, cette comparaison n'étant réalisée que sur un seul exemple, ce n'est pas suffisant pour comparer l'efficacité relative des deux classes de méthodes (coalescence et distance).

Enfin, aussi bien l'étude de Morris *et al.* (2002) que les comparaisons sur le jeu de données CFTR ont été réalisées dans le cas de modèles où un locus de susceptibilité a un effet majeur (La mutation $\Delta F508$ est responsable de 66% des cas de mucoviscidose). Il serait

[7] L'évolution indépendante des haplotypes depuis un ancêtre commun (généalogie en étoile) n'est pas réaliste et conduit à considérer plusieurs fois certains évènements de mutation ou de recombinaison ce qui fausse l'estimation de la localisation. Les différents auteurs qui utilisent ce modèle lui applique alors une correction

[8] Une généalogie en plusieurs morceaux permet de faire l'hypothèse que plusieurs mutations indépendantes se sont produites au locus de susceptibilité et de prendre en compte la présence éventuelle de cas sporadiques (individus dont la maladie est due à un autre facteur génétique ou environnemental)

Méthode	Position[a]	Variabilité
McPeek et Strahs (1999)	950	440-1460 (intervalle de confiance)
Morris *et al.* (2000)	800	610-1070 (intervalle de probabilité)
Liu *et al.* (2001)	–	820-930 (intervalle de probabilité)
Morris *et al.* (2002)	850	650-1000 (intervalle de probabilité)
Molitor *et al.* (2003a)	880	620-670,770-970[b](intervalle de probabilité)
Molitor *et al.* (2003b)	869,8	–
Zöllner et Pritchard (2005)	867	814-920 (intervalle de probabilité)

La mutation $\Delta F508$ qui est responsable de 66% des cas de mucoviscidose est située à la position 885

[a] La position est donnée en kilobase (kb) et est mesurée depuis l'extrémité gauche de la région CFTR

[b] La distribution observée est bimodale

TAB. 1.3: Localisation de la mutation impliquée dans la susceptibilité à la mucoviscidose

intéressant d'étudier le comportement de ces méthodes pour des modèles plus complexes (pénétrance incomplète, phénocopies, plusieurs locus de susceptibilité, etc.). De plus, il faut noter que ces méthodes tirent leur information des recombinaisons et considèrent les mutations comme des paramètres de nuisances. Lorsque le locus de susceptibilité est situé dans un bloc haplotypique (i.e. où il ne s'est pas produit de recombinaisons), ces méthodes ne permettent pas de localiser finement les locus de susceptibilité.

1.1.2.3 Tests multiples et catégories rares

Les tests multiples Le problème des tests multiples se pose très souvent pour les études d'association. Il est très important de réaliser un bonne correction qui limite le nombre de faux positifs mais sans entraîner une trop grande baisse de puissance.

La correction de Bonferroni Il s'agit de la correction la plus simple. Si on veut évaluer la significativité de n tests pour un risque de première espèce global de α, il faudra considérer un risque individuel pour chaque test égal à $\frac{\alpha}{n}$.

Cette correction est bien adaptée dans le cas de tests indépendants, sinon, elle est trop conservatrice et entraîne une perte de puissance. Il est alors plus judicieux (mais plus long en temps de calcul) d'utiliser une procédure de permutation.

La permutation simple La procédure de permutation simple ne permet pas, en elle-même, de corriger pour les tests multiples. Elle permet plutôt d'estimer les p-values correspondant à une valeur de la statistique lorsque sa distribution n'est pas connue. C'est, par exemple, le cas pour la statistique du χ^2 : lorsqu'on a des petits échantillons, elle ne suit pas exactement la loi du χ^2. Nous exposerons rapidement le principe de la simple permutation car elle permet de mieux comprendre ensuite les autres procédures de permutation.

Pour évaluer la p-value d'une statistique S_1 par une procédure de permutation classique dans le cas d'une analyse cas/témoins, n réplicats sont générés en redistribuant aléatoirement les statuts maladie des individus, ce qui revient à affecter des statuts maladie aux indi-

vidus indépendamment de leur génotype. En calculant les valeurs s_i de la statistique utilisée sur ces réplicats, on obtient une distribution de cette statistique sous l'hypothèse nulle d'absence d'association (H_0). Les s_i sont alors classés par ordre décroissant et on peut obtenir le rang r de S_1 dans cette distribution. La p-value associée à S_1 est calculée ainsi :

$$p = \frac{r}{n}$$

Comme la précision de la valeur de p dépend de n, il faut réaliser un nombre suffisant de permutations (10 000 à 100 000).

La permutation de Westfall et Young (1993) La procédure décrite par Westfall et Young (1993) permet d'obtenir des p-values pour k tests en tenant compte de la multiplicité des tests réalisés. Tout comme la procédure de permutation simple, elle nécessite la constitution de n réplicats en redistribuant les statuts maladie des individus. Le principe de cette permutation de Westfall et Young (1993) est décrit sur la Figure 1.3 à partir d'un exemple (4 tests).

	Test 1	Test 2	Test 3	Test 4
Statistique observée	12,8	8,6	7,5	5,7
Réplicat 1	8,3	7,3	10,5	4,8
	↓	↓	↓	↓
	r_1	r_2	r_3	r_4

Les valeurs de la statistique sont calculées pour les 4 tests et sont classées par ordre décroissant. Pour faciliter les explications, les tests sont renumérotés dans l'ordre des valeurs de la statistique. Les valeurs de la statistique sont calculées pour les n réplicats. Du fait des permutations, les valeurs de la statistique ne sont pas nécessairement ordonnées d'un réplicat à l'autre, de la même façon que les valeurs de la statistique observées.

On définit alors un compteur r_i pour chaque test i, $1 \leq i \leq 4$. Pour chaque réplicat et pour chaque test, on va alors rechercher si ce compteur doit être incrémenté ou non : si la valeur de la statistique observée pour le test i est supérieure à la valeur maximale de la statistique obtenue pour les tests i à 4 du réplicat considéré, alors le compteur est incrémenté de 1. Par exemple, pour le 1er réplicat, on a :

- test 1 : 12,8 > max(8,3 ; 7,3 ; 10,5 ; 4,8), donc on incrémente r_1 de 1 ;
- test 2 : 8,6 $\leq$ max(7,3 ; 10,5 ; 4,8), donc on n'incrémente pas r_2 ;
- test 3 : 7,5 $\leq$ max(10,5 ; 4,8) donc on n'incrémente pas r_3 ;
- test 4 : 5,7 > 4,8, donc on incrémente r_4 de 1.

On agit de même pour les $n - 1$ réplicats restants. On obtient finalement 4 valeurs de r_i qui permettent de calculer 4 valeurs de $p_i* = \frac{r_i}{n}$.

Or, les statistiques étant classées par ordre décroissant, on doit observer une croissance des p-values corrigées p_i. Elles sont donc finalement calculées ainsi :

$$p_1 = p_1*, \quad p_2 = max(p_1, p_2*), \quad p_3 = max(p_2, p_3*) \quad p_4 = max(p_3, p_4*)$$

FIG. 1.3: La permutation de Westfall et Young (1993)

La permutation de Ge *et al.* (2003) Cette procédure permet aussi d'obtenir des p-values pour k tests en tenant compte de la multiplicité des tests réalisés. Elle a été développée par Ge *et al.* (2003) dans le cadre de l'analyse de puces à ADN puis, a été appliquée en génétique épidémiologique par Becker et Knapp (2004b). Son principe est assez différent. Il est basé sur le fait que lorsque plusieurs tests sont réalisés, on souhaite conclure à une association entre le gène et la maladie si le test ayant la plus petite p-value non corrigée (p^a_{min}) a une p-value suffisamment faible. Le principe de la permutation de Ge *et al.* (2003) consiste à obtenir une distribution des p_{min} sous l'hypothèse nulle d'absence d'association et à évaluer la p-value en tenant compte des tests multiples en recherchant le rang de p^a_{min} dans la distribution des p_{min} sous H_0. Il est détaillé sur la Figure 1.4 à partir d'un exemple.

Les catégories rares Nous avons vu précédemment que pour tester l'association entre un gène candidat et une maladie, il peut être intéressant d'utiliser l'information conjointe de plusieurs marqueurs. Or, lorsqu'un grand nombre de marqueurs est pris en compte, on va obtenir un grand nombre de catégories haplotypiques ou génotypiques. Dans ce cas, les méthodes haplotypiques et les méthodes génotypiques sont confrontées aux mêmes problèmes :

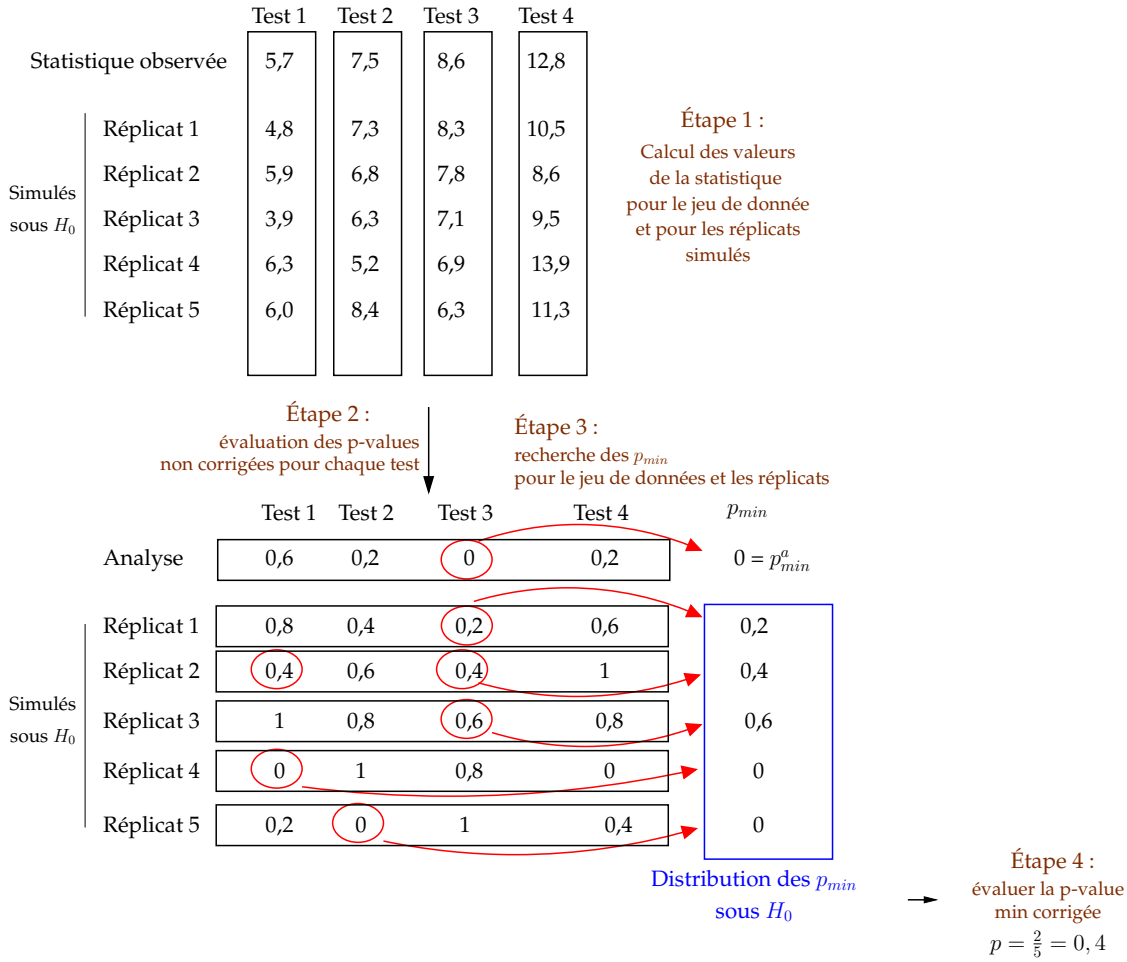
- Nombre de degrés de liberté élevé, ce qui va entraîner une diminution de puissance des tests de distribution haplotypique ou génotypique ;
- Présence de catégories à effectif réduit, ce qui pose des problèmes de petits échantillons pour la réalisation des tests (pour les χ^2 en particulier) ;
- Augmentation du nombre de tests pour les stratégies de recherche d'haplotypes partagés ou de tests de l'effet d'une catégorie par rapport aux autres.

Pour éviter la perte de puissance due au grand nombre de catégories et aux petits échantillons, une stratégie intéressante consiste à regrouper les catégories. Il existe plusieurs stratégies de regroupement de catégories. Nous allons décrire les principales dans le cas de tests haplotypiques, mais elles sont généralement adaptables pour des tests génotypiques.

La stratégie la plus simple consiste à regrouper tous les haplotypes rares en une seule catégorie (Jannot *et al.*, 2004; Tzeng *et al.*, 2003). Si on a R haplotypes fréquents, alors l'analyse comportera $R+1$ catégories. Cette méthode présente toutefois l'inconvénient de former une catégorie très hétérogène, ce qui ne facilite pas l'interprétation des résultats.

Une seconde stratégie également décrite par Tzeng *et al.* (2003) consiste à regrouper les haplotypes rares avec les haplotypes fréquents dont ils ne sont séparés que par un seul pas mutationnels. Lorsqu'un haplotype rare est séparé de plusieurs haplotypes fréquents par un seul pas mutationnel, il est regroupé avec l'haplotype le plus fréquent. Les haplotypes rares qui sont séparés des haplotypes fréquents par plusieurs pas mutationnel sont regroupés en une catégorie à part. Tzeng *et al.* (2003) ont comparé les deux stratégies décrites ci-dessus et ils ont conclu que, dans le cas d'haplotypes formés de peu de marqueurs (3), elles ont des performances comparables. Lorsque les haplotypes sont formés d'un plus grand nombre de marqueurs (6), la seconde stratégie est plus puissante.

Ces deux stratégies nécessitent de définir un critère permettant de distinguer les haplotypes rares et les haplotypes fréquents. Ce critère peut être absolu comme dans l'étude de Jannot *et al.* (2003) (les catégories comprenant moins de 5 haplotypes (ou diplotypes) sont groupées ensemble) ou relatif comme dans l'étude de Tzeng *et al.* (2003) (les haplotypes dont la fréquence est supérieure à c fois celle de l'haplotype le plus fréquent sont considérés comme fréquents. Dans leur étude, Tzeng *et al.* (2003) ont déterminé par simulations que 0,2 est une valeur optimale pour c).



Étape 1 : calcul des valeurs de la statistique pour les 4 tests sur le jeu de données observées et sur les 5 réplicats simulés.

Étape 2 : évaluation des p-values non corrigées pour chaque test. Par exemple, pour calculer la p-value associée à la valeur du test 1 du réplicat 1 (4,8), on compte le nombre R de réplicats (analyse du jeu de données comprise) pour lesquelles la statistique a une valeur supérieure ou égale à 4,8. Dans cet exemple, $R = 4$. la p-value non corrigée est obtenue en divisant R par le nombre total de réplicats, sans compter l'analyse du jeu de données (ici, $p = \frac{4}{5} = 0,80$). On calcule ainsi les p-values associées à toutes les valeurs de la statistique. On obtient alors le tableau du bas ;

Étape 3 : pour chaque réplicat (analyse du jeu de données comprise), on recherche la p-value la plus faible parmi les p-values des 4 tests ;

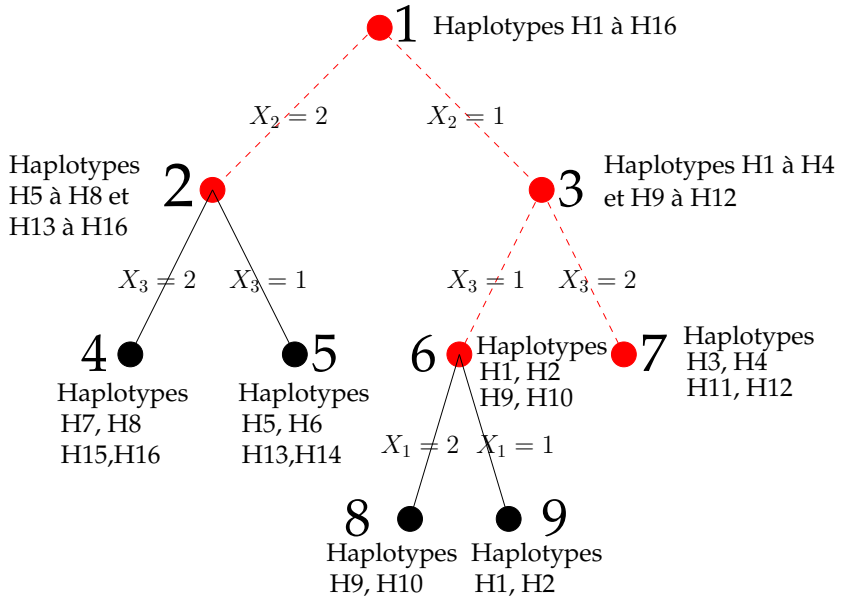
Étape 4 : on dispose alors d'une distribution des p-values minimales (p_{min}) sous l'hypothèse nulle d'absence d'association ainsi que d'une valeur p^a_{min} pour le jeu de données à analyser. On détermine la p-value corrigée en comptant le nombre de réplicats pour lesquels la p-value minimale a une valeur inférieure ou égale à p^a_{min}, et en divisant ce nombre par le nombre total de réplicats. Ici, $p^a_{min} = \frac{2}{5} = 0,4$.

FIG. 1.4: La permutation de Ge *et al.* (2003)

Une troisième stratégie propose de regrouper les haplotypes rares avec la catégorie haplotypique qui leur est la plus similaire. La similarité haplotypique peut être définie de différentes manières. Jannot *et al.* (2004) proposent, par exemple, d'utiliser le nombre de SNPs identiques entre les deux haplotypes comme mesure de leur similarité. Jannot *et al.* (2004) ont montré que cette stratégie était plus puissante que le simple regroupement des haplotypes rares en une seule catégorie.

Un quatrième type de stratégie consiste à regrouper les haplotypes en utilisant des arbres de classification. L'ensemble des données est placé à la racine de l'arbre de classification. Il est alors divisé en deux groupes, de manière à ce que les deux nœuds fils constituent les groupes les plus homogènes possibles. Chaque nœud fils peut ensuite être divisé à son tour en deux, la procédure se répétant récursivement jusqu'à ce que chaque catégorie d'haplotypes forme un nœud terminal de l'arbre. Le critère utilisé pour former des groupes homogènes peut varier. Par exemple, Yu *et al.* (2005) proposent d'utiliser une mesure basée sur la longueur de la région où les haplotypes possèdent les mêmes allèles, comme définie par Bourgain *et al.* (2000). La Figure 1.5 représente un arbre de classification obtenu à partir d'un jeu de données contenant 16 haplotypes formés de 4 SNPs.

Haplotypes	
H1	1111
H2	1112
H3	1121
H4	1122
H5	1211
H6	1212
H7	1221
H8	1222
H9	2111
H10	2112
H11	2121
H12	2122
H13	2211
H14	2212
H15	2221
H16	2222



Les X_i représentent les allèles du SNPs i. La classification est faite sur le critère de longueur de la région où les haplotypes partagent les même allèles (Bourgain *et al.*, 2000).

Sur cet arbre de classification, Yu *et al.* (2005) définissent des sous-arbres : chacun d'eux contient la racine et une partie des nœuds internes (la structure en pointillé est l'un des sous-arbres possibles). Le niveau d'un sous-arbre est donné par le nombre de groupes terminaux qu'il contient (le sous-arbre en pointillé est de niveau 3).

FIG. 1.5: Regroupement de catégories en utilisant un arbre de classification d'après Yu *et al.* (2005)

Yu *et al.* (2005) ont proposé de définir des groupes (ou *clusters*) d'haplotypes correspondant à l'ensemble des nœuds terminaux d'un sous-arbre. Pour un arbre de classification

donné, il y a donc de nombreuses façons de définir les groupes. Yu *et al.* (2005) proposent de sélectionner certains regroupements d'haplotypes sur la base d'un score attribué au sous-arbre qui les définit et de ne réaliser des tests d'association que sur ces regroupements. Par exemple, si le sous-arbre en pointillé de la Figure 1.5 est sélectionné, le test d'association effectué concernera les groupes d'haplotypes 2, 6 et 7. La p-value globale du test correspond alors à la plus petite p-value obtenue parmi tous les tests d'association effectués sur les sous-arbres, en réalisant une procédure de permutation pour tenir compte des tests multiples. Ils ont montré par simulations que, pour de nombreux modèles génétiques, cette méthode est très souvent plus puissante que la méthode de regroupement de Tzeng *et al.* (2003) (voir ci dessus) et s'avère toujours beaucoup plus puissante qu'un simple test haplotypique sans regroupement.

Enfin, une dernière stratégie possible consiste à regrouper les haplotypes en fonction de leur histoire évolutive. Elle a été proposée pour la première fois par Templeton *et al.* (1987). L'idée de base de cette stratégie est qu'une mutation responsable de la susceptibilité à une maladie se retrouvera sur tous les descendants de l'haplotype initialement muté. Les haplotypes porteurs de l'allèle à risque devraient donc former des clades et être de ce fait identifiables sur un arbre racontant l'histoire évolutive des haplotypes (arbre phylogénétique). En outre, Templeton *et al.* (1987) ont aussi montré que l'utilisation de la phylogénie des haplotypes pouvait permettre de restreindre la région dans laquelle se trouve le site de susceptibilité à la maladie. À partir des travaux de Templeton *et al.* (1987), d'autres tests basés sur l'histoire évolutive des haplotypes ont été développés. Dans la suite de notre travail, nous nous intéresserons plus en détail à ces tests. Mais, avant de les décrire, il nous parait utile de rappeler quelques notions fondamentales de phylogénie.

1.2 La phylogénie

1.2.1 Les concepts de base

1.2.1.1 Définition

Une phylogénie peut être définie comme une représentation de l'*histoire évolutive* des objets étudiés. Ces objets — ou taxons — peuvent être de nature variée. Traditionnellement, il s'agit d'espèces, mais cela peut aussi être des genres, des familles, des populations, des individus, des gènes, voire des unités non biologiques telles que les langues.

Les taxons sont définis par des caractères, c'est-à-dire par des traits observables les décrivant. Ces caractères peuvent être de différents types : morphologique (couleur des yeux, par exemple) ou moléculaire (position dans une séquence nucléotidique ou d'acides aminés). Ces caractères peuvent exister sous différentes formes ou *états*. On dit alors que le caractère est polymorphe.

Dans la suite de ce travail, les taxons que nous considérerons sont des haplotypes d'individus malades ou témoins. Ils sont constitués de caractères binaires (les SNPs) dont les états seront notés 0 et 1.

1.2.1.2 Les arbres phylogénétiques

Vocabulaire utile pour décrire les arbres phylogénétiques Dès l'émergence du concept dans les années 1860, les phylogénies sont représentées sous la forme d'arbres montrant les

relations de parenté entre les taxons. La description des arbres phylogénétiques fait appel à un vocabulaire particulier. Nous ne rappellerons ici que les termes essentiels pour la suite de ce manuscrit.

Les relations entre les taxons sont ici représentées sur un graphe appelé en phylogénie arbre non enraciné. Ses sommets externes représentent les taxons terminaux (A, B, C, D et E sur la Figure 1.6) et sont appelés les *feuilles*. Les sommets internes (1, 2 et 3 sur la Figure 1.6) sont appelés des *nœuds*. Les segments qui relient deux sommets (nœuds ou feuilles) sont appelés les *branches*.

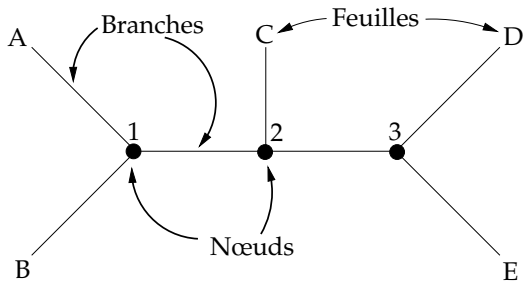
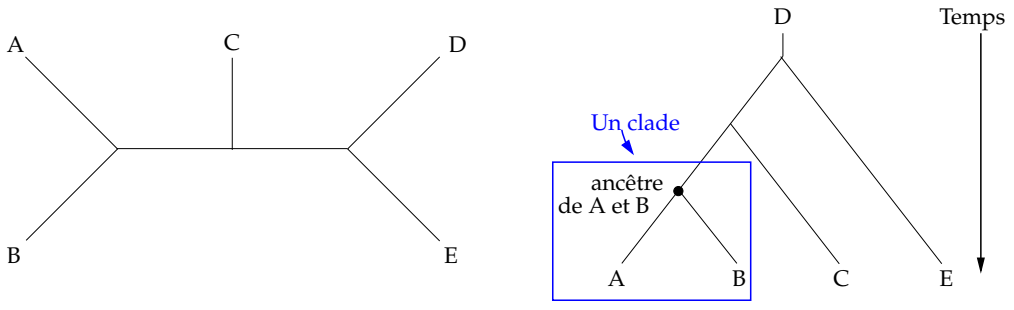


FIG. 1.6: Description d'un arbre phylogénétique non enraciné

La notion d'ancêtre et d'enracinement Un arbre non enraciné tel que celui représenté à la Figure 1.6 n'est pas une phylogénie car il ne raconte pas l'histoire évolutive des taxons. Pour devenir une phylogénie, un arbre doit être enraciné, c'est-à-dire qu'on doit définir un point de départ — une racine — et un sens de lecture correspondant au sens d'écoulement du temps. Lorsque l'arbre est enraciné, un nœud correspond alors à l'ancêtre des feuilles qui en dérivent et un ensemble de taxons constitué d'un ancêtre et de tous ses descendants forme un *clade*. La Figure 1.7 présente les topologies d'un arbre non enraciné (Figure 1.7a) et du même arbre enraciné en considérant le taxon D comme racine (Figure 1.7b).



(a) Arbre non enraciné

(b) Arbre enraciné

FIG. 1.7: Différences entre un arbre enraciné par un ancêtre et un arbre non enraciné

Il existe plusieurs méthodes pour enraciner les arbres. Parmi elles, la plus classique consiste à ajouter un ou plusieurs taxons n'appartenant pas au groupe étudié. Par exemple,

notre étude concernant des haplotypes humains, nous pourrions prendre comme groupe externe — ou *extra-groupe* — des haplotypes de singes. Comme ces séquences ont divergé de l'ancêtre commun avant la divergence des séquences du groupe interne (i.e. les haplotypes humains), il serait possible d'orienter l'arbre. Cette procédure est représentée à la Figure 1.8. Il faut bien noter ici que l'extra-groupe ne correspond pas à l'ancêtre commun à toutes les séquences et qui se trouve au niveau de la racine.

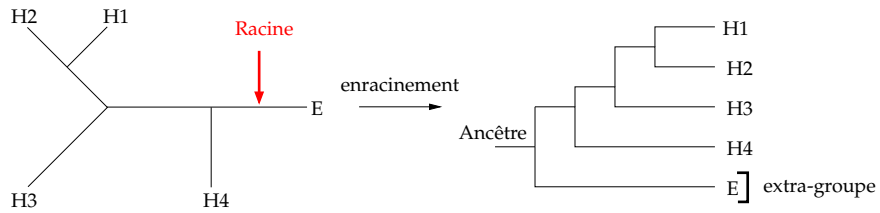


FIG. 1.8: Enracinement sur un groupe externe
Soient 4 haplotypes humains (H1 à H4) formant le groupe étudié. Pour enraciner l'arbre, on ajoute à ces données un haplotype de singe (E). L'arbre est alors orienté de telle manière que l'haplotype de singe soit celui qui diverge le premier de l'ancêtre commun à l'homme et au singe.

Une autre méthode d'enracinement consiste simplement à proposer une séquence ancestrale. Dans le cas d'haplotypes de SNPs qui sera traité par la suite, deux propositions sont faites, basées sur l'idée qu'une fréquence d'un allèle ou d'un haplotype est un signe d'ancienneté :

- La séquence la plus fréquente dans l'échantillon est supposée être ancestrale. On la notera par la suite H_{maj} ;
- La séquence consensus formée des allèles les plus fréquents à chaque marqueur est supposée être ancestrale. On la notera par la suite H_{cons}.

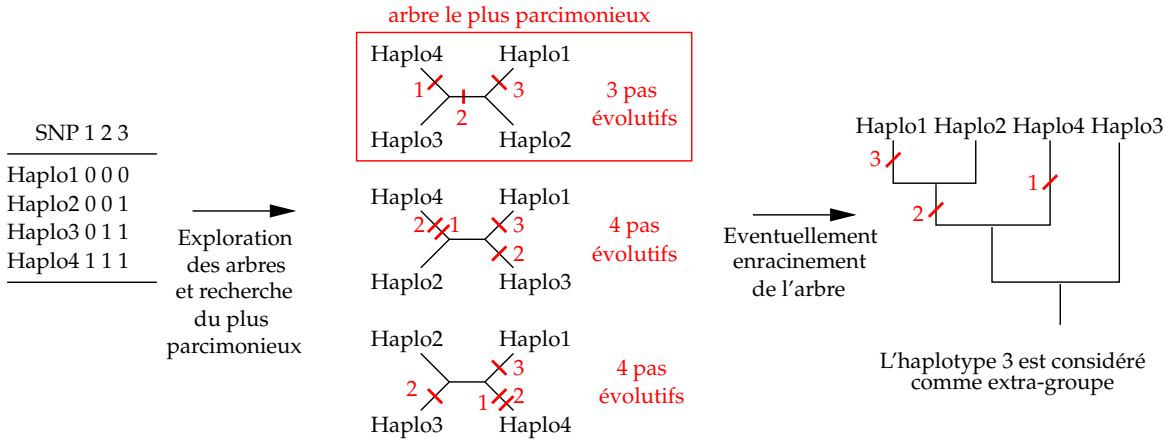
1.2.2 Les méthodes de reconstruction phylogénétique

Il existe trois types de méthodes de reconstruction phylogénétique : les méthodes de parcimonie, les méthodes de distance et les méthodes de maximum de vraisemblance. Dans la suite de ce paragraphe, nous allons décrire brièvement les principes de ces trois méthodes de reconstruction phylogénétique en nous focalisant sur les notions importantes pour la suite de notre travail. Pour une étude plus exhaustive des méthodes de reconstruction phylogénétique, on pourra se référer aux ouvrages de Darlu et Tassy (1993) et de Felsenstein (2004a).

1.2.2.1 Les méthodes de parcimonie

Les méthodes de parcimonie sont basées sur l'analyse indépendante des caractères. Elles reposent sur l'idée générale que l'arbre qui reflète le mieux l'histoire évolutive des taxons est celui qui minimise le nombre d'évènements évolutifs subis par ces caractères.

Principe Les données analysées par parcimonie sont des ensembles de caractères définis pour plusieurs taxons. Dans le cas d'haplotypes de SNPs, l'analyse par parcimonie est réalisée selon la procédure décrite sur la Figure 1.9.



Étape 1 : À partir de 4 haplotypes, 3 arbres différents sont possibles. Pour chacun de ces arbres, on cherche le nombre de changements d'état des SNPs — on parle souvent de pas évolutifs — nécessaire pour rendre compte de la distribution des SNPs observée dans les différents haplotypes. L'arbre retenu est celui qui requiert le plus petit nombre de pas évolutifs (principe de parcimonie). Sur la figure, les changements d'état des SNPs 1, 2 et 3 sont indiqués par un petit trait sur les branches concernées ;

Le nombre d'arbres possibles augmente très rapidement avec le nombre de taxons. Par exemple, pour N taxons, il existe $\frac{(2N-3)!}{2^{N-1}(N-1)!}$ arbres binaires[a] enracinés possibles(Felsenstein, 2004b). Des heuristiques sont donc souvent utilisées pour optimiser les chances de trouver l'arbre le plus parcimonieux en un temps de calcul raisonnable.

Étape 2 : Une fois l'arbre le plus parcimonieux identifié, il peut être enraciné. Sur la figure, nous avons choisi l'haplotype 3 comme extra-groupe.

FIG. 1.9: L'analyse par parcimonie

[a]les arbres binaires sont des arbres pour lesquels il y a toujours strictement deux descendants pour un ancêtre

Lorsque l'arbre est enraciné, on peut identifier les états des caractères qui définissent les différents clades de l'arbre. Ils seront appelés *synapomorphies* (états dérivés, partagés par tous les taxons appartenant à un clade).

L'homoplasie Lorsque deux taxons possèdent un caractère sous le même état, il peut être soit hérité d'un ancêtre commun — il s'agit alors d'une homologie — soit non hérité d'un ancêtre commun et il s'agit alors d'une *homoplasie*. On peut définir deux types d'homoplasie :

- *La convergence :* il s'agit de l'apparition indépendante chez plusieurs taxons d'un même état du caractère (Figure 1.10a) ;
- *La réversion :* il s'agit de l'apparition chez un ou plusieurs taxons, d'un caractère ayant l'état ancestral (Figure 1.10b).

Sur la Figure 1.10, on remarque que les deux configurations, convergence ou réversion, sont également parcimonieuses. Lorsque les changements d'état des caractères sont inférés sur l'arbre, il n'y aura donc pas toujours une solution unique. Certains logiciels de reconstruction phylogénétique comme PAUP permettent de favoriser les convergences (option

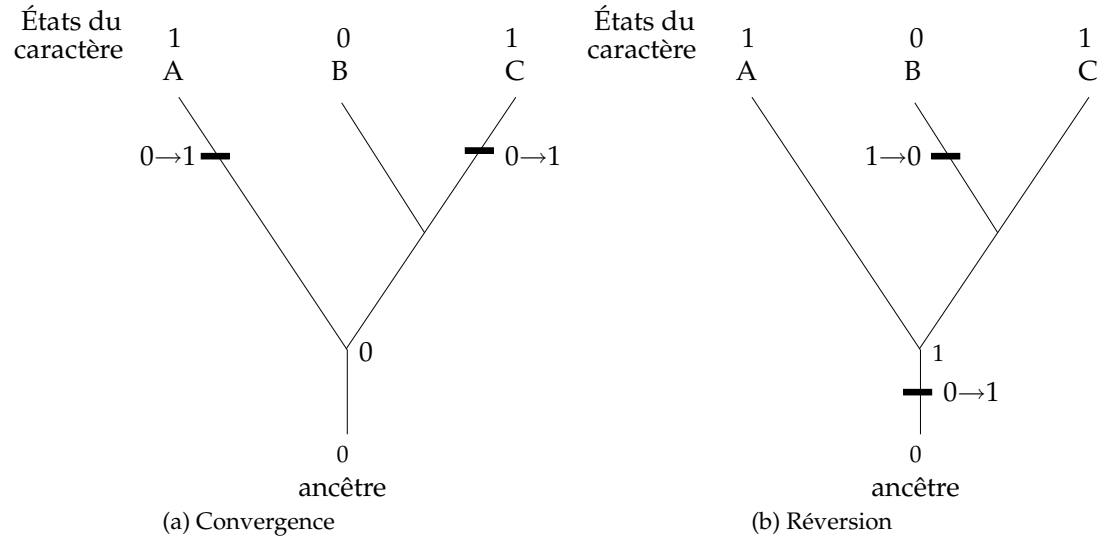


FIG. 1.10: Notion de convergence et de réversion

`deltran` — *delayed transformation* — car les changements d'état se produisent plus tardivement dans l'arbre) ou les réversions (option `acctran` — *accelerated transformation* — car les changements d'état se produisent plus tôt dans l'arbre).

En cladistique, les relations de parenté étant fondées sur les synapomorphies, l'homoplasie correspond à du "bruit" qui brouille l'information phylogénétique et que les méthodes de parcimonie cherchent à minimiser. La mesure du degré d'homoplasie permet alors de mesurer la quantité d'information phylogénétique contenue dans les données analysées. Cette mesure se fait classiquement en utilisant l'indice de cohérence (CI, *consistency index*). Le CI correspond au rapport entre le nombre minimal de transformations nécessaires pour expliquer les états de tous les caractères — noté R — et le nombre effectif de transformations dans l'arbre[9]. Ainsi, lorsque le nombre d'homoplasies diminue, la valeur du CI tend vers 1. La valeur minimum du CI n'est pas zéro. Elle a été estimée à environ 0,2 en simulant des données pour lesquelles les états des caractères sont distribués aléatoirement chez les différents taxons. Par conséquent, des arbres ayant un CI inférieur à 0,3 contiendront particulièrement peu d'information phylogénétique et seront donc à examiner d'un œil critique.

Les algorithmes de reconstruction phylogénétique

Les algorithmes exacts Les algorithmes exacts garantissent de trouver la solution optimale. Ils peuvent le faire par une recherche exhaustive, c'est-à-dire par l'évaluation de toutes les topologies possibles. Cependant, lorsque le nombre de taxons augmente, le nombre d'arbres devient rapidement si élevé qu'une telle recherche exhaustive n'est pas réalisable dans un temps de calcul raisonnable.

[9] Pour un caractère c donné possédant s états, le nombre minimal de transformations nécessaires pour rendre compte de tous ses états est $r_c = s - 1$. R est donc égal à la somme des r_c pour tous les caractères.

L'algorithme de *branch and bound* (Hendy et Penny, 1982) garantit la solution optimale pour des données comprenant jusqu'à 20 ou 30 taxons sans toutefois réaliser une recherche exhaustive. Le principe de cet algorithme est d'utiliser un arbre de référence, régulièrement mis à jour, pour limiter l'exploration des différentes topologies.

Les algorithmes heuristiques Lorsque le nombre de taxons dépasse 20 ou 30, seuls les algorithmes heuristiques permettent d'obtenir une solution en un temps raisonnable. Ils ne garantissent cependant pas l'obtention de la solution optimale.

La plupart des algorithmes heuristiques sont actuellement basés sur un algorithme d'addition pas à pas (*stepwise addition*) : l'algorithme de Wagner (Kluge et Farris, 1969; Farris, 1970) que les auteurs des différents logiciels de phylogénie ont amélioré. Le principe de cet algorithme est d'ajouter un à un les taxons en minimisant à chaque fois le nombre de transformations nécessaires dans l'arbre.

Les procédures de réarrangement Les algorithmes heuristiques ne donnent pas forcément l'arbre le plus parcimonieux. De plus, les algorithmes d'addition pas à pas sont sensibles à l'ordre d'introduction des séquences. Sur les arbres retenus par une heuristique, des procédures de réarrangement de branches (*branch swapping*) sont donc réalisées pour maximiser les chances de trouver l'arbre ou les arbres les plus parcimonieux. Dans la suite de notre travail, nous avons utilisé l'algorithme de réarrangement de branches nommé TBR (*tree bisection and reconnection*)(Swofford et Olsen, 1990). Il s'agit d'une procédure de réarrangement global : l'arbre initial est découpé en sous-arbres qui sont ensuite reconnectés ensemble par chacune de leurs branches. Le détail de la procédure est expliqué sur la Figure 1.11.

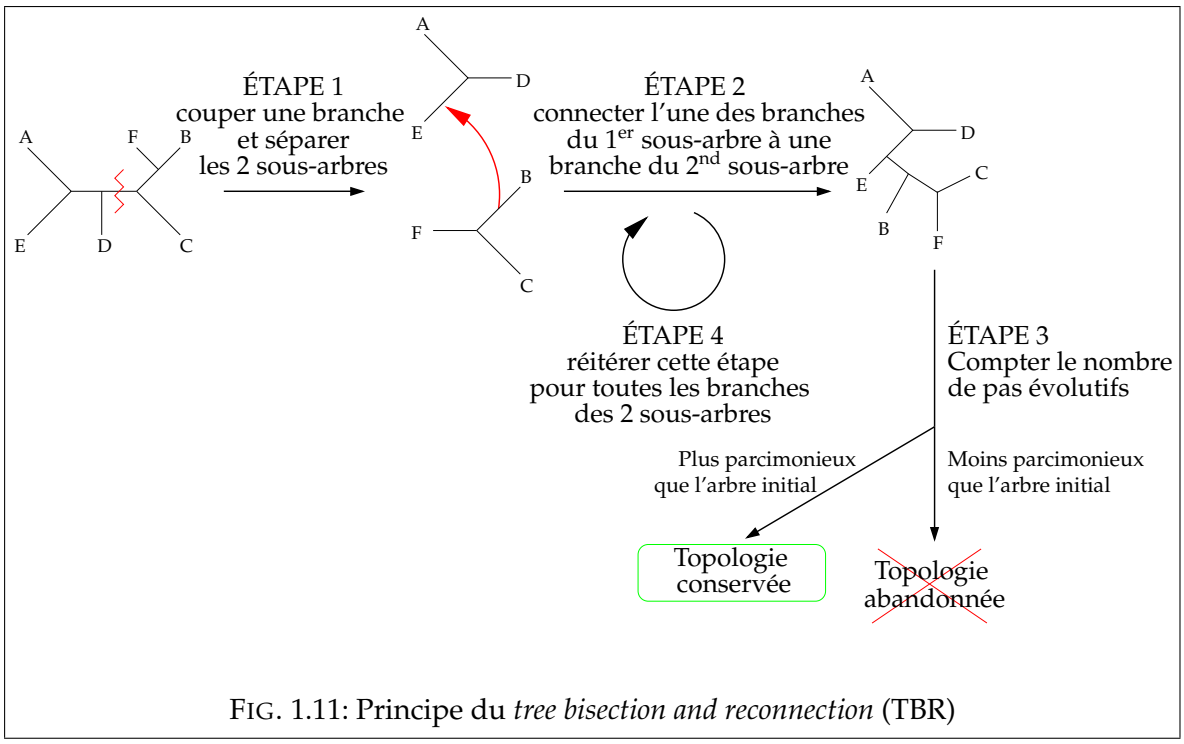


FIG. 1.11: Principe du *tree bisection and reconnection* (TBR)

Les algorithmes de réarrangement de branches permettent de bien améliorer l'efficacité des heuristiques, mais ils augmentent le temps de calcul. Enfin, si plusieurs arbres de longueur minimale existent, ils ne permettent pas toujours de les obtenir tous.

Les logiciels de reconstruction phylogénétique utilisant une méthode de parcimonie Il existe plusieurs logiciels de reconstruction phylogénétique ayant implémenté une méthode de parcimonie. Dans la suite de notre travail, nous utiliserons principalement les logiciels `PAUP` (*Phylogenetic Analysis Using Parsimony*) (Swofford, 2002) et `PHYLIP` (*Phylogeny Inference Package*) (Felsenstein, 2004c). Ces logiciels sont les plus couramment utilisés pour réaliser des reconstructions phylogénétiques. L'une des différences[10] entre les deux logiciels, importante pour la suite de notre travail, est que les deux options `acctran` et `deltran` ne sont pas implémentées dans `PHYLIP` alors qu'elles le sont dans `PAUP` . `PAUP` résout donc les ambiguïtés aux nœuds selon l'option choisie par l'utilisateur alors que `PHYLIP` leur affecte l'état " ?" (donnée indéterminée).

1.2.2.2 Les méthodes phénétiques

Les approches phénétiques, aussi appelées méthodes de distance, ont été développées dans les années 1960. Elles sont basées sur l'utilisation d'une mesure de similitude globale entre les taxons étudiés. Pour ces méthodes, plus les taxons se ressemblent globalement, plus leurs liens de parenté sont étroits. Comme elles ne considèrent pas les caractères individuellement, ces méthodes ne permettront pas de déterminer les états ancestraux des caractères aux différents nœuds de l'arbre.

Principe Les données utilisées pour une analyse de distance consistent en une matrice contenant les distances entre les taxons deux à deux. La mise en œuvre d'une méthode de distance nécessite deux étapes :

- Établissement de la matrice de distance : il faut pour cela choisir la mesure de distance appropriée au traitement des données ;
- Estimation de l'arbre à partir de cette matrice.

La notion de distance La notion de distance est liée à celle de similitude : plus la similitude entre deux taxons est forte, plus la distance entre eux est faible. Les distances vérifient trois propriétés :

Soit i, j et k trois taxons. On notera d_{ij}, la distance entre i et j.

- si $i = j$ alors $d_{ij} = 0$ (la distance d'un taxon à lui-même est nulle)
- si $i \neq j$ alors $d_{ij} > 0$ (la distance est positive)
- $d_{ij} = d_{ji}$ (la distance est commutative)

Il existe de nombreux indices de distance qui vont permettre d'établir la matrice des distances. Pour les haplotypes de SNPs, on peut utiliser les indices développés pour les séquences de nucléotides ou d'acides aminés où l'on ne prend en compte que l'identité ou la différence des états de caractère. Par exemple, un des indices les plus simples est l'indice de concordance de Sokal et Michener (1958). Il s'agit d'un indice de similitude correspondant à la proportion de caractères qui sont dans le même état chez les deux taxons étudiés. Pour effectuer la transformation de cet indice de similitude s en distance d, il suffit d'effectuer

[10] Hormis le prix : `PHYLIP` est gratuit alors que `PAUP` est sous licence payante !

la transformation $d = 1 - s$. Plusieurs autres indices ont été décrits. Pour une revue de ces indices, on pourra se reporter à la section 2 du chapitre VII du livre de Darlu et Tassy (1993), et à l'ouvrage de Sokal et Sneath (1963)

Les algorithmes de reconstruction d'arbres Il existe de nombreux types d'algorithmes de reconstruction phénétique. Nous n'exposerons que les deux algorithmes qui seront utiles dans la suite de notre travail : les méthodes de classification hiérarchique combinatoire et la méthode de neighbor-joining (Saitou et Nei, 1987).

Les méthodes de classification hiérarchique combinatoire Ce type de méthode n'est pas le plus utilisé car il fait l'hypothèse que les taux de mutations sont les mêmes sur toutes les branches (hypothèse d'une *horloge moléculaire*). Si cette hypothèse n'est pas vérifiée, la structure de l'arbre et les longueurs des branches risquent d'être fausses.

Le principe de ces méthodes est le suivant :

- Regroupement des deux groupes i et j les plus ressemblants (*i.e.* entre lesquels la distance est la plus petite) en une unité évolutive hypothétique ij qui formera un nœud de l'arbre ;
- Calcul d'une nouvelle matrice de distance dans laquelle ij remplace i et j. Le mode de calcul diffère selon les méthodes. Pour la méthode d'UPGMA (*Unweighted pair-group method using arithmetic averages*) (Sokal et Michener, 1958), la distance $d_{ij,k}$ du groupe ij au groupe k est calculée ainsi :

$$d_{ij,k} = (\frac{n_i}{n_i + n_j})d_{i,k} + \frac{n_j}{n_i + n_j})d_{j,k}$$

  où n_i et n_j correspondent au nombre de taxons déjà agglomérés dans les groupes i et j
- Réitérer les deux étapes précédentes jusqu'à ce que tous les taxons aient été agglomérés en une seule unité évolutive hypothétique.

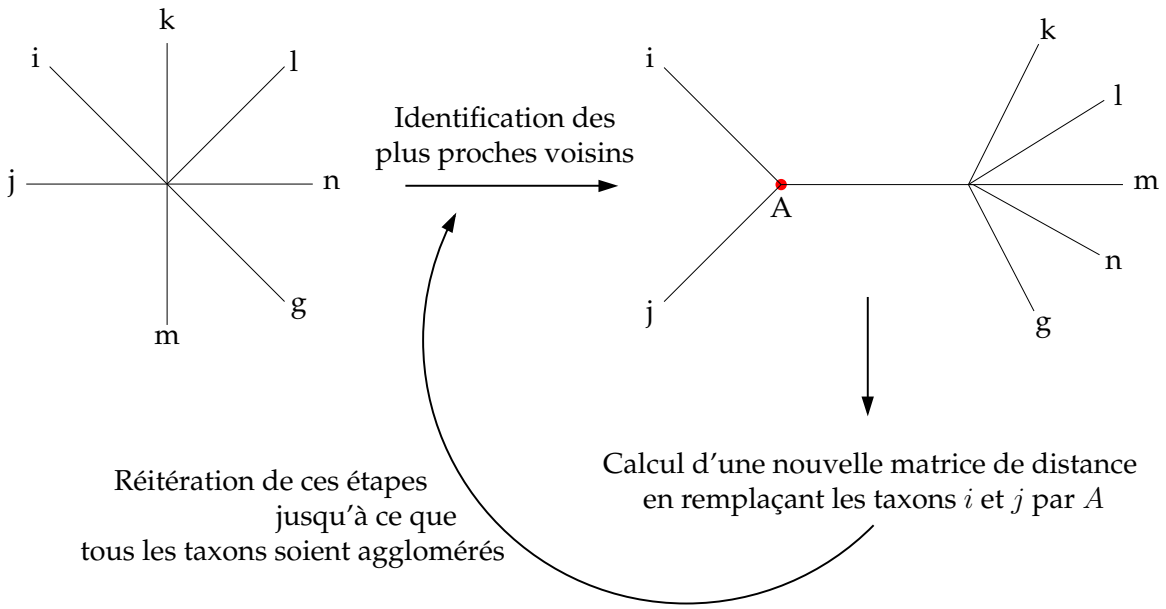
La méthode du neighbor-joining (NJ) C'est l'une des méthode de distance les plus utilisées. Contrairement aux méthodes de classification hiérarchique combinatoire, elle n'impose pas l'hypothèse d'une horloge moléculaire. Le détail de l'algorithme est représenté sur la Figure 1.12.

1.2.2.3 La méthode de maximum de vraisemblance

Cette méthode nécessite la spécification d'un modèle d'évolution constitué d'un ensemble de paramètres θ sur lesquels on peut formuler différentes hypothèses. Les paramètres du modèle d'évolution sont à la fois les probabilités d'évènements évolutifs (probabilités de changements d'état des caractères, probabilité des différents états, etc.) et la topologie (i.e. la forme et les longueurs de branches) de l'arbre.

La méthode de maximum de vraisemblance (notée classiquement ML pour le terme anglais *maximum likelihood*) s'appuie sur un ensemble de données D, ici organisées en une matrice taxons $\times$ caractères. L'ensemble des arbres possibles va alors être parcouru et on va estimer les valeurs des paramètres du modèle évolutif qui maximisent la vraisemblance[11] de l'arbre. L'arbre finalement retenu sera celui pour lequel la valeur de la vraisemblance est la plus élevée.

[11] La vraisemblance de l'arbre est proportionnelle à la probabilité des données sachant l'arbre.



Considérons le cas de n taxons. L'algorithme démarre à partir d'une phylogénie en étoile.

Étape 1 : regroupement des deux taxons qui sont "les plus proches voisins" parmi les N taxons considérés. Pour chaque regroupement d'une paire de taxons i, j, on calcule la somme totale des distances $S_{i,j}$. Les taxons i et j les plus proches voisins seront ceux qui minimisent $S_{i,j}$. $S_{i,j}$ est calculé à partir de la matrice des distances : $S_{i,j} = D_{i,j} - u_i - u_j$, où

- $D_{i,j}$ est la distance observée entre les taxons i et j
- u_i (resp. u_j) $= \sum_{i \neq j}^{n} \frac{D_{i,j}}{n-2}$

Les longueurs des branches $e_{i,A}$ et $e_{j,A}$ se calculent alors ainsi :

$$e_{i,A} = \frac{1}{2} D_{i,j} + \frac{1}{2}(u_i - u_j)$$

Étape 2 : Calcul d'une nouvelle matrice de distance en remplaçant i et j par A. La distance de A aux autres taxons est calculée ainsi :

$$d_{A,k} = \frac{1}{2} \times (d_{i,k} + d_{j,k} - d_{i,j})$$

Sur la nouvelle matrice, on recherche à nouveau les deux taxons qui sont les plus proches voisins, et ainsi de suite jusqu'à ce que tous les taxons soient agglomérés et que toutes les longueurs de branches aient été calculées.

FIG. 1.12: L'algorithme du neighbor-joining (NJ)

Calcul de la vraisemblance d'un arbre. Un modèle d'évolution doit tout d'abord être défini. Dans le cas des SNPs :

- les probabilités de changement d'état $0 \to 1$ et $1 \to 0$ sont supposées égales : $P(0 \to 1) = P(1 \to 0) = \rho$;
- les probabilités des états 0 et 1 chez l'ancêtre sont respectivement f et $1 - f$;
- on suppose que tous les SNPs évoluent à la même vitesse.

La vraisemblance est calculée en faisant deux hypothèses :

- évolution indépendante des différents caractères[12] ;
- évolution indépendante dans les différentes branches.

À titre d'exemple, intéressons-nous au calcul de la vraisemblance de l'arbre T[13] représenté sur la Figure 1.13.

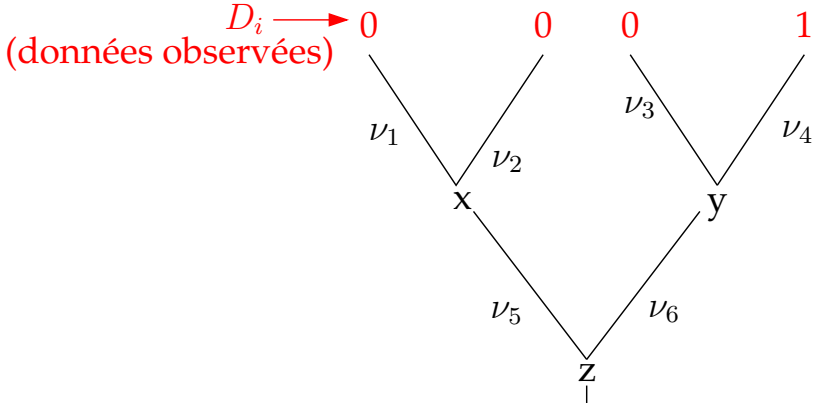


FIG. 1.13: Exemple d'arbre utilisé pour décrire le calcul de la vraisemblance
Les états du caractère i sont indiqués à chaque nœud. Aux nœuds externes, les états sont 0 ou 1. Aux nœuds internes, les états étant inconnus, on les note x, y et z. Les longueurs de branches sont notées ν_i.

Soit le caractère i dont les états aux nœuds sont représentés sur la Figure 1.13. La probabilité des données de ce caractère i (notées D_i) sur l'arbre T est la somme des probabilités de toutes les distributions possibles des états 0 et 1 aux nœuds internes :

$$P(D_i|T) = \sum_x \sum_y \sum_z P(0,0,0,1,x,y,z|T)$$

Une fois la vraisemblance calculée pour chacun des n caractères, l'hypothèse d'indépendance des caractères permet de calculer la vraisemblance de l'arbre :

$$L(T) = P(D|T) = \prod_{i=1}^{n} P(D_i|T)$$

[12] Cette hypothèse est aussi vraie pour les méthodes de parcimonie.

[13] L'arbre T est représenté enraciné, mais, comme le modèle d'évolution est réversible (c'est-à-dire que les changements d'état sont aussi probables dans un sens que dans l'autre), le choix du point d'enracinement n'a pas d'influence sur la vraisemblance calculée. De fait, la vraisemblance calculée correspond à celle de l'arbre non enraciné.

Stratégies de recherche des topologies d'arbres La recherche des différentes topologies se fait par le même genre d'algorithmes que dans le cas de l'analyse par parcimonie, c'est-à-dire par *Branch and bound* pour une recherche exhaustive, ou par des heuristiques suivies de réarrangements (TBR ou NNI).

Optimisation des caractères Par maximum de vraisemblance, on ne peut pas inférer les états de caractères aux différents nœuds autrement qu'en terme de probabilités pour une topologie d'arbre donnée où les longueurs des branches $\nu_1, \nu_2, \ldots, \nu_n$ sont connues.

Soit θ, l'ensemble des longueurs de branches $\nu_1, \nu_2, \ldots, \nu_n$. Soit x l'ensemble des états du caractère i pour tous les nœuds terminaux et y l'ensemble des états du caractère i pour tous les nœuds ancestraux conjointement. On peut alors estimer y en maximisant la probabilité d'observer y connaissant x et θ (Yang *et al.*, 1995) :

$$P(y|x,\theta) = \frac{P(y)P(x|y,\theta)}{P(x,\theta)}$$

On peut aussi estimer la probabilité *a posteriori* de l'état d'un caractère pour un nœud donné k (Yang *et al.*, 1995). Soit y_k l'état du caractère i au nœud k.

$$P(y_k = 0|x,\theta) = \frac{\sum_{y,y_k=0} P(y)P(x|y,\theta)}{P(x,\theta)}$$

correspond à la probabilité pour le caractère d'être à l'état 0 au nœud k. On peut alors déterminer l'état le plus probable pour ce caractère et la probabilité qui lui est associée. Ce sont ces probabilités que nous utiliserons pour la suite de notre étude.

Les logiciels de reconstruction phylogénétique utilisant des méthodes de maximum de vraisemblance Les logiciels `PAUP` et `PHYLIP` , précédemment décrits dans le paragraphe concernant les méthodes de parcimonie (voir page 29) permettent aussi de reconstruire un arbre en utilisant une méthode de maximum de vraisemblance. Cependant, les algorithmes implémentés dans ces logiciels sont extrêmement lents et ne permettent pas de réaliser des simulations. Pour la suite de notre travail, nous avons donc choisi d'utiliser le logiciel `PHYML` (Guindon et Gascuel, 2003) qui est très rapide (quelques secondes de calcul pour obtenir un arbre à partir de 24 haplotypes formés de 20 SNPs sur un pentium III, 930 MHz, 512 Mo de RAM). Ce logiciel ne permet pas d'obtenir les probabilités des états des caractères aux nœuds. Nous avons donc choisi d'utiliser le logiciel `PAML` (Yang, 1997) pour obtenir les probabilités des états de caractères pour les nœuds de l'arbre reconstruit par `PHYML` . `PAML` aurait pu permettre de rechercher l'arbre phylogénétique, mais son algorithme est lent et pas optimal au dire même de son auteur. Il n'a donc été utilisé que pour l'optimisation des caractères.

1.2.2.4 Les avantages et inconvénients de chaque méthode

L'intérêt principal des méthodes de distance est leur rapidité : en un temps relativement court, il est possible d'obtenir des phylogénies pour un grand nombre de taxons. De même, comme c'est une similitude globale qui est calculée, le nombre de caractères dans les séquences ne va pas augmenter le temps de calcul. Il existe deux principaux problèmes aux

méthodes de distances. Tout d'abord, le calcul d'une distance globale entraîne une perte d'information par rapport à une analyse basée sur les caractères dans la mesure où il ne sera pas possible d'inférer les états des différents caractères aux nœuds. Enfin, les distances estimées sur l'arbre peuvent poser des problèmes d'interprétation, en particulier dans le cas où elles sont négatives.

Les méthodes de parcimonie permettent en revanche de connaître les états de caractères aux différents nœuds de l'arbre ce qui permet une interprétation facile des arbres en terme d'histoire évolutive des caractères. Cependant, il faut noter que cette reconstruction des états de caractères est parfois ambiguë car il est souvent difficile de choisir entre favoriser les convergences ou les réversions. Les méthodes de parcimonie possèdent aussi l'inconvénient d'être inconsistantes, c'est-à-dire que lorsque la quantité de données augmente, elles ne vont pas nécessairement converger vers l'arbre vrai. Ce phénomène s'observe en particulier dans le cadre du problème de l'attraction des longues branches[14]. Enfin, les méthodes de parcimonie ne peuvent évidemment donner des reconstructions correctes que dans le cas où leur postulat de base — à savoir que l'évolution est parcimonieuse — est vérifié.

Un avantage des méthodes de maximum de vraisemblance est qu'elles permettent de tester des hypothèses phylogénétiques. En effet, il est facile de calculer la vraisemblance des données sous deux modèles emboîtés différents puis de les tester en utilisant un test de rapport de vraisemblance. Jusqu'en 2003, le principal inconvénient de ces méthodes était le temps de calcul très long dès que le nombre de taxons dépassait une dizaine. Depuis 2003, Guindon et Gascuel (2003) ont développé le logiciel `PHYML` qui permet de résoudre ce problème. Les méthodes de maximum de vraisemblance nécessitent aussi la spécification explicite d'un modèle. Cela peut être considéré comme un avantage car les hypothèses sont ainsi toutes posées, ou comme une faiblesse car, si le modèle est mal choisi, cela peut avoir des conséquences sur l'arbre retenu.

[14] Lorsque deux taxons ont accumulé beaucoup de mutations, ils vont avoir tendance à être très proches dans l'arbre reconstruit par parcimonie

Chapitre 2

L'utilisation des phylogénies d'haplotype en génétique épidémiologique

Sommaire

2.1 L'analyse des clades emboîtés (Templeton *et al.*, 1987) 37
2.1.1 Descriptif de la méthode 37
2.1.1.1 Reconstruction phylogénétique 37
2.1.1.2 Analyse de l'arbre 38
2.1.2 Étude de puissance 41
2.2 La méthode EHAP (Seltman *et al.*, 2003) 41
2.2.1 Descriptif de la méthode 41
2.2.1.1 Reconstruction phylogénétique 41
2.2.1.2 Analyse de l'arbre 41
2.2.2 Étude de puissance 42
2.2.2.1 Résultats 42
2.3 La méthode CLADHC (Durrant *et al.*, 2004) 44
2.3.1 Descriptif de la méthode 45
2.3.1.1 Reconstruction phylogénétique 45
2.3.1.2 Analyse de l'arbre 45
2.3.2 Étude de puissance 46
2.3.2.1 Simulation des données 47
2.3.2.2 Résultats 47
2.3.3 Localisation du site de susceptibilité 48
2.4 La méthode de Tree Scanning (Templeton *et al.*, 2005) 49
2.4.1 Description de la méthode 49
2.4.1.1 Reconstruction phylogénétique 49
2.4.1.2 Analyse de l'arbre 49
2.4.2 Étude de puissance 50
2.5 La méthode de Tzeng (Tzeng, 2005) 51
2.5.1 Descriptif de la méthode 51
2.5.2 Étude de puissance 52
2.5.2.1 Méthode de simulation 52
2.5.2.2 Résultats 52
2.5.2.3 Analyse de données 53

2.6 Utilisation des méthodes phylogénétiques sur des données 53
2.6.1 Reconstruction de l'arbre 53
2.6.2 Type d'analyse statistique réalisée 54
2.6.3 Détection de l'association 54
2.6.4 Localisation des sites de susceptibilité 55
2.7 Bilan sur les méthodes d'association et de localisation de sites de susceptibilité basées sur des phylogénies 59

Préambule

Dans le premier chapitre de cette thèse, nous avons défini une phylogénie comme une représentation de l'histoire évolutive. Il en découle qu'une phylogénie est nécessairement représentée par un arbre enraciné, avec un sens de lecture bien défini. Or, dans la suite de ce travail — et plus particulièrement dans ce chapitre — le terme "phylogénie" sera employé dans un sens très *large, incluant les arbres non enracinés et même les réseaux comportant des réticulations. De la même façon, toutes les méthodes d'association décrites dans ce chapitre seront qualifiées de "phylogénétiques" ou "basées sur la reconstruction de l'histoire évolutive", même dans le cas où aucune phylogénie (au sens strict) n'est reconstruite. Cette grossière approximation est due au fait que nous avons conservé la plupart des termes employés par les auteurs des différentes méthodes dans la littérature. Or, ces auteurs n'utilisent pas la terminologie exacte et ont systématiquement qualifié leur méthode de "phylogénétique", voire, même généralement, de "cladistique"*[1]*, quelle que soit la méthode de reconstruction et le type d'arbre (enraciné ou non) analysé. Nous espérons que le lecteur ainsi averti ne nous tiendra pas rigueur de cet abus de langage.*

Dans le chapitre précédent, nous avons vu que, pour tester l'association entre un gène candidat et une maladie, il peut être intéressant d'utiliser l'information conjointe des marqueurs situés sur un même chromosome, en particulier en utilisant l'information haplotypique. Cependant, lorsque le nombre de marqueurs considérés augmente, l'intérêt de ces méthodes peut beaucoup diminuer, surtout s'il y a peu de déséquilibre de liaison entre les marqueurs. En effet, dans ce cas, le nombre d'haplotypes présents dans l'échantillon peut être très grand, ce qui entraîne deux problèmes : le nombre de degrés de liberté du test d'association va aussi être élevé et certains haplotypes vont être très rares, ce qui va poser des problèmes de petits échantillons lors de la réalisation du test d'association. La puissance des tests haplotypiques risque donc d'être beaucoup diminuée. Comme nous l'avons vu au chapitre précédent, une solution intéressante consiste à regrouper les haplotypes rares selon des critères *a priori*. Templeton *et al.* (1987) sont les premiers à avoir proposé de réaliser les groupements en fonction de l'histoire évolutive des haplotypes et à avoir développé un test d'association utilisant l'arbre phylogénétique. Ils ont aussi montré que l'utilisation de la phylogénie des haplotypes peut permettre de restreindre la région dans laquelle se trouve le site

[1] Afin de ne pas trop embrouiller les phylogénéticiens débutants et ne pas trop choquer les phylogénéticiens confirmés qui pourraient lire ce manuscrit, nous avons évité l'emploi du terme "analyse cladistique" qui est pourtant le plus utilisé pour qualifier ces méthodes...

de susceptibilité à la maladie. Par la suite, plusieurs autres tests basés sur des phylogénies ont été développés.

Dans ce chapitre, nous détaillerons ces différents tests. Pour chacun d'entre eux, nous exposerons son principe et l'étude de sa puissance lorsqu'elle a été réalisée. Enfin, nous terminerons ce chapitre par une revue des différentes applications de ces méthodes sur divers jeux de données.

2.1 L'analyse des clades emboîtés (Templeton *et al.*, 1987)

Entre 1987 et 1995, A.R. Templeton a développé la première méthode phylogénétique permettant de détecter l'association entre un gène et une maladie. Il l'a appelée analyse des clades emboîtés (*nested clade analysis*)[2] et l'a décrite dans une série de 5 articles (Templeton *et al.*, 1987, 1988, 1992; Templeton et Sing, 1993; Templeton, 1995). Parmi ces articles, trois concernent strictement la méthode : les deux premiers articles (Templeton *et al.*, 1987, 1988) décrivent la méthode appliquée à des données quantitatives, respectivement pour des populations haploïdes ou pseudo-haploïdes (populations formées d'homozygotes, étude du chromosome Y, etc.) et diploïdes. Le dernier article (Templeton, 1995) propose une extension de la méthode sur des données cas/témoins. Les deux autres articles (Templeton *et al.*, 1992; Templeton et Sing, 1993) sont plus techniques et traitent de la manière de construire et d'estimer la validité d'un cladogramme lors d'une analyse phylogénétique et des problèmes posés par la recombinaison.

2.1.1 Descriptif de la méthode

La méthode de Templeton suppose que les haplotypes ont été préalablement reconstruits. Les marqueurs qui forment les haplotypes doivent être biallélіques (à l'origine, Templeton a développé la méthode pour des marqueurs correspondant à la présence ou à l'absence de sites de restriction) et la recombinaison entre les marqueurs doit être négligeable afin que la phylogénie reconstruite n'en soit pas affectée.

2.1.1.1 Reconstruction phylogénétique

Dans ses premiers articles, Templeton préconise de reconstruire la phylogénie des haplotypes en utilisant les algorithmes de parcimonie implémentés dans PAUP (Swofford, 2002) et dans PHYLIP (Felsenstein, 2004c). Par la suite, il a développé son propre algorithme adapté au traitement de données intra-spécifiques (Templeton *et al.*, 1992). Cet algorithme a été implémenté récemment par Clement *et al.* (2000) dans le programme TCS . Son principe est de travailler successivement sur les haplotypes qui diffèrent entre eux par 1, 2, ...,j ,..., n mutations. À chaque niveau de mutation j, et pour toutes les paires d'haplotypes concernées, on estime la probabilité P_j que les relations entre les deux haplotypes soient parcimonieuses, c'est-à-dire que chaque différence observée entre ces haplotypes soit due à un seul changement d'état et non pas à des mutations récurrentes à ce locus. Si P_j est supérieur à un

[2] Le terme "clade" est ici impropre car Templeton *et al.* (1987) reconstruisent en fait un réseau. Pour la description de la méthode, nous préférerons utiliser le terme de groupe. Cependant, nous conserverons le nom "analyse des clades emboîtés" que Templeton *et al.* (1987) ont donné à leur méthode.

seuil (fixé à 0.95 par Templeton *et al.*), les haplotypes sont alors connectés entre eux. L'algorithme s'arrête soit lorsque tous les haplotypes forment un seul réseau, soit lorsque les haplotypes constituent plusieurs réseaux non chevauchants dont les connections entre eux ont toutes des probabilités d'être parcimonieuses P_j inférieures au seuil. Dans ce dernier cas, l'algorithme réunit entre eux les différents réseaux par les haplotypes entre lesquels les proportions de sites non parcimonieux sont minimales.

Avec l'algorithme de Templeton, on obtient au final non pas des arbres, mais des réseaux, avec des boucles traduisant des ambiguïtés dans la résolution de la phylogénie. Ces boucles peuvent être soit conservées pour l'analyse des clades emboîtée, soit cassées selon les règles définies par Crandall et Templeton (1993). Les branches sont cassées de manière à ce que les haplotypes les plus fréquents — donc les plus anciens d'après la théorie de la coalescence — constituent des nœuds internes de l'arbre phylogénétique et aient un plus grand nombre de descendants (en terme d'haplotypes différents) que les haplotypes rares. Selon les marqueurs analysés, d'autres règles permettent de réduire les ambiguïtés. Par exemple, pour les RFLPs (*Restriction Length Fragment Polymorphisms*), la perte d'un site de restriction est considérée comme plus probable que la création d'un nouveau site de restriction.

2.1.1.2 Analyse de l'arbre

Définition de groupes emboîtés Le principe de l'analyse consiste à regrouper les haplotypes dans des groupes de plus en plus grands jusqu'à ce que tous les haplotypes soient inclus dans un unique groupe. Les groupes formés sont de deux types : les groupes dits "externes" qui ne sont reliés qu'à un seul autre groupe et les groupes dits "internes" qui sont reliés à plusieurs groupes. Les haplotypes constituent les groupes de niveau 0. Pour former les groupes de niveau supérieur $(i+1)$, le principe est de regrouper les groupes externes de niveau i avec les groupes internes de niveau i dont ils sont séparés par un seul pas évolutif. De plus, chaque groupe de niveau i ne peut appartenir qu'à un seul groupe de niveau $i+1$. Au sein d'un groupe de niveau $i+1$, les différents haplotypes sont donc séparés par au maximum $i+1$ pas évolutifs. La Figure 2.1 illustre ce processus.

Nous n'avons présenté ici que la méthode de définition des groupes emboîtés sur un arbre (i.e. sans boucles liées aux ambiguïtés). Templeton a développé des règles permettant de traiter le cas des réseaux (Templeton et Sing, 1993) que nous ne développerons pas ici.

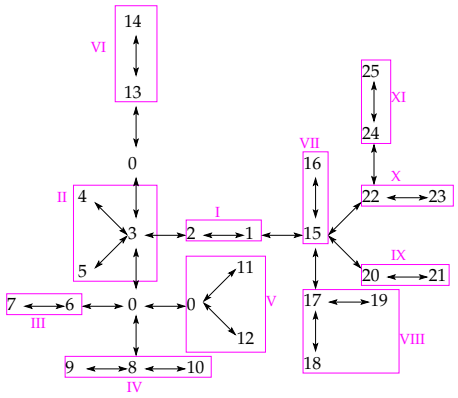
Analyse statistique Le principe est de réaliser une analyse emboîtée des données sur l'arbre. Le processus est décrit sur la Figure 2.1 dans le cas d'une analyse de données cas/témoins. Il consiste à réaliser à chaque niveau un test d'homogénéité de la distribution des malades et des témoins dans les différents groupes.

Imaginons que sur la Figure 2.1, le test du niveau 2 soit significatif. Tous les contrastes deux à deux entre groupes adjacents dans l'arbre sont alors réalisés. Dans notre exemple, les contrastes entre les groupes adjacents α et γ, α et β, α et δ, δ et ϵ seront réalisés afin de localiser sur l'arbre les groupes entre lesquels une hétérogénéité phénotypique existe.

Pour l'analyse de données quantitatives, le principe est le même, mais Templeton propose d'utiliser une NANOVA (Templeton *et al.*, 1987) (*Nested ANalysis Of VAriance*)

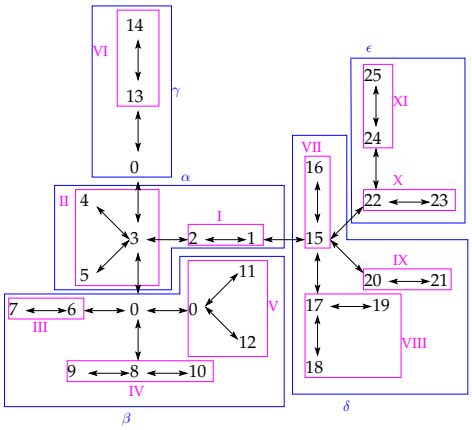
Localisation du site de susceptibilité Templeton *et al.* (1987) ont aussi proposé une méthode permettant de préciser la région dans laquelle est situé le variant fonctionnel impli-

Au niveau 0, les 25 haplotypes sont considérés séparément. On compare la distribution des haplotypes chez les malades et chez les témoins en réalisant un test de χ^2 à 24 degrés de liberté.
Les haplotypes externes sont ensuite regroupés avec l'haplotype interne dont ils sont séparés par un seul pas évolutif pour former les groupes de niveau 1.



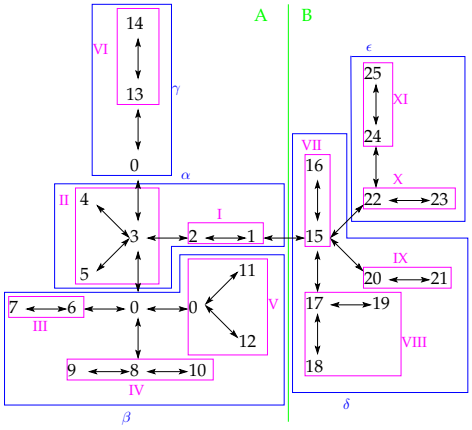
Au niveau 1, Il existe onze groupes de niveau 1 (numérotés de I à XI sur la figure, en rose). Les haplotypes des malades et des témoins sont répartis dans ces groupes. Pour chaque groupe, on calcule les nombres attendus d'haplotypes malades et témoins. On réalise alors un test de χ^2 à 10 degrés de liberté comparant les distributions des haplotypes malades et témoins dans ces 11 groupes.

Les groupes externes de niveau 1 sont alors regroupés avec le groupe interne de niveau 1 dont ils sont séparés par un seul pas évolutif pour former les groupes de niveau 2



Au niveau 2, Il existe cinq groupes de niveau 2 (nommés α, β, γ, δ et ϵ, en bleu). De la même manière qu'au niveau 1, on réalise un test de χ^2 à 4 degrés de liberté comparant les distributions des haplotypes malades et témoins dans ces 5 groupes.

Les groupes externes de niveau 2 sont alors regroupés avec le groupe interne de niveau 2 dont ils sont séparés par un seul pas évolutif pour former les groupes de niveau 3.



Au niveau 3, Il existe deux groupes de niveau 3 (nommés A et B, en vert). De la même manière qu'au niveau 2, on réalise un test de χ^2 à 1 degrés de liberté comparant les distributions des haplotypes malades et témoins dans ces 2 groupes.

Les deux groupes de niveau 3 peuvent être regroupés en un groupe de niveau 4 qui contient alors tous les haplotypes. De ce fait, l'analyse s'arrête.

FIG. 2.1: L'analyse des groupes emboîtés de Templeton

Les numéros correspondent aux haplotypes. Lorsque la valeur 0 apparaît, il s'agit d'un haplotype qui n'a pas été observé dans l'échantillon mais qui est nécessaire pour la reconstruction de l'arbre.

qué dans la maladie. Cette méthode est basée sur l'étude d'haplotypes recombinants. Son principe est représenté sur la Figure 2.2.

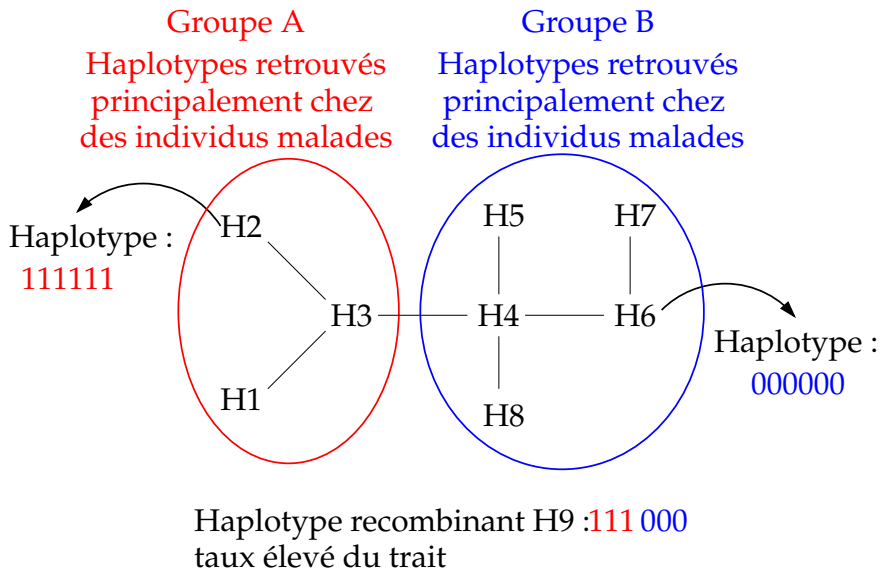


FIG. 2.2: Localisation du site de susceptibilité en utilisant les haplotypes recombinants

Soit l'haplotype H9, issu de la recombinaison des haplotypes H2 et H6. On le retrouve principalement chez des individus malades. Son effet phénotypique ressemble donc à celui de l'haplotype H2. On peut alors penser que le variant fonctionnel est situé sur la région chromosomique que H9 a héritée de H2, c'est-à-dire, dans cet exemple, avant le quatrième marqueur.

Cette méthode ne peut s'appliquer qu'à trois conditions :

1. s'il est possible d'identifier les haplotypes recombinants et leurs parents sans ambiguïté. En pratique, les évènements de recombinaison étant supposés rares, on postule que les haplotypes recombinants sont des haplotypes rares et qu'ils sont issus de la recombinaison de deux haplotypes fréquents. Deux problèmes se posent alors :
   - Les haplotypes rares ne sont pas forcément tous des recombinants ;
   - Lorsque les haplotypes sont constitués de marqueurs biallélIques, on peut souvent identifier plusieurs paires de parents potentiels pouvant générer par recombinaison un haplotype rare. Le problème se pose encore plus quand il y a des données manquantes.
2. s'il est possible d'attribuer une valeur phénotypique à l'haplotype recombinant. En pratique, comme nous l'avons souligné au point précédent, les haplotypes recombinants sont souvent rares. Pour les études cas/témoins, il est alors difficile de dire si cet haplotype est porté par une majorité de malades ou de témoins. Cette méthode pourra donc surtout être employée dans le cas d'études quantitatives ;
3. s'il existe des haplotypes recombinants issus de la recombinaison de deux haplotypes appartenant à des groupes qui diffèrent significativement pour la valeur du trait étudié. Ces haplotypes ne sont alors pas utilisés pour la construction du cladogramme, ni pour l'analyse emboîtée.

En définitive, cette méthode ne peut que très rarement s'appliquer, surtout quand le nombre d'haplotypes dans l'échantillon est élevé.

2.1.2 Étude de puissance

La méthode de Templeton a été appliquée sur plusieurs jeux de données (voir page 53) soit pour analyser des traits quantitatifs (voir Tableau 2.1) soit pour analyser des données qualitatives (voir Tableau 2.2). En revanche, sa puissance n'a jamais été étudiée par simulations.

2.2 La méthode EHAP (Seltman *et al.*, 2003)

Seltman *et al.* se sont tout d'abord intéressés aux méthodes phylogénétiques dans le cadre des études familiales. En 2001, ils ont développé le ET-TDT (*Evolutionary Tree — Transmission Disequilibrium Test*) (Seltman *et al.*, 2001), un test qui associe la méthodologie du TDT haplotypique[3] et de l'analyse phylogénétique. Par la suite, Seltman *et al.* ont étendu la méthode du ET-TDT à l'analyse de données cas/témoins. Ils ont alors diffusé leur test sous la forme d'un logiciel nommé EHAP (*Evolutionary-based Haplotype Analysis Package*) (Seltman *et al.*, 2003). Dans la suite de cette description, nous désignerons la méthode développée par Seltman *et al.* sous le nom de EHAP , quel que soit le type de données analysées (données familiales ou données en population)

2.2.1 Descriptif de la méthode

Contrairement aux autres méthodes d'analyse basées sur des phylogénies, EHAP ne nécessite pas que les haplotypes soient préalablement reconstruits. Il utilise donc les données génotypiques et incorpore l'incertitude de reconstruction des haplotypes dans l'analyse statistique de l'arbre en tenant compte de toutes les configurations haplotypiques possibles pour chaque individu et de leurs probabilités respectives.

2.2.1.1 Reconstruction phylogénétique

Dans EHAP , Seltman *et al.* (2003) ont implémenté leur propre méthode de reconstruction phylogénétique. Pour l'ensemble des haplotypes, une matrice de distances entre haplotypes est construite. La distance utilisée correspond au nombre de pas évolutifs séparant les haplotypes. Les haplotypes séparés par un seul pas évolutif dans la matrice sont ensuite reliés pour former un réseau. Enfin, à partir du réseau, un arbre phylogénétique est obtenu en cassant les boucles selon les principes établis par Crandall et Templeton (1993) (voir page 37).

2.2.1.2 Analyse de l'arbre

Principe La méthode EHAP est basé sur l'utilisation de modèles de régression où la probabilité d'être atteint est exprimée en fonction des risques e^{β_i} associés à chaque haplotype ou groupe d'haplotypes i. L'arbre des haplotypes est utilisé pour définir des groupes emboîtés selon le même principe que Templeton *et al.* (1987) et pour tester par un test de rapport de vraisemblance, à la fois

[3] voir page 11 (principe du TDT) et page 15 (TDT haplotypique)

– si les risques des différentes entités (haplotypes ou groupes d'haplotypes) du groupe sont significativement différents. Si ce n'est pas le cas (test non significatif), alors un seul paramètre β est utilisé au niveau supérieur pour représenter les risques associés à l'ensemble du groupe ;
– si les risques sont égaux à 1, ce qui permet de tester l'association avec la maladie.

Le détail de l'algorithme est illustré à l'aide d'un exemple sur la Figure 2.3.

Correction pour les tests multiples Afin d'éviter la détection de fausses associations dues aux multiples tests à un degré de liberté réalisés, Seltman *et al.* proposent soit d'utiliser une correction de Bonferroni (option par défaut de EHAP), soit d'évaluer une p-value corrigée par une procédure de permutations du même type que la permutation de Ge *et al.* (2003) décrite à la page 20. La procédure de permutations est particulièrement conseillée lorsqu'il y a des haplotypes rares dans l'arbre phylogénétique.

2.2.2 Étude de puissance

Seltman *et al.* (2001) ont étudié la puissance de EHAP dans le cas de données familiales. Ils ont simulé 400 trios[4] dont la moitié sont génotypés pour les deux parents et l'autre moitié pour un seul parent, à partir des 11 haplotypes de l'étude de Hallman *et al.* (1994) sur l'apolipoprotéine B. Ils ont testé 13 modèles différents :

– tous les haplotypes dans un groupe associé à la maladie ont soit le même risque, soit des risques différents ;
– les haplotypes associés à la maladie appartiennent soit tous au même groupe, soit à des groupes différents ;
– certains modèles supposent de la recombinaison ;
– certains modèles supposent la présence d'erreurs dans la reconstruction de l'arbre phylogénétique.

La puissance du test est évaluée sur 200 réplicats et elle est comparée à celle de deux tests TDT hapotypiques[5] classiques réalisés avec le programme TRANSMIT (Clayton, 1999) :

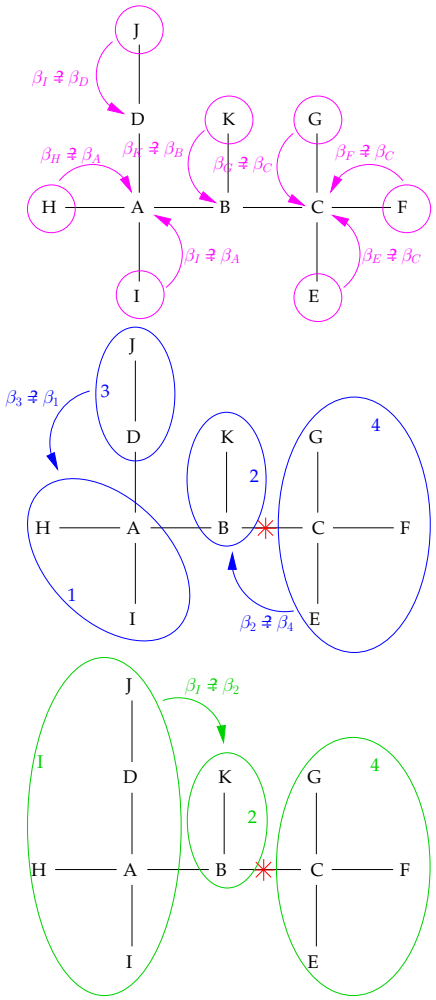
– TDT_{all} qui teste conjointement l'effet de tous les haplotypes
– TDT_{max} qui teste séparément l'effet de chaque haplotype versus l'effet de tous les autres

Afin d'éviter les problèmes de petits échantillons, en TDT classique, ils n'ont pris en compte que les haplotypes dont la fréquence est supérieure à 0.04. Pour le EHAP , ils ont rassemblé les haplotypes rares (fréquence < 0.14) avec l'haplotype fréquent le plus proche dans l'arbre phylogénétique.

2.2.2.1 Résultats

Comparaison des trois méthodes pour la détection de l'association La comparaison des deux TDT montre que le TDT_{max} est bien moins puissant que le TDT_{all}, en particulier lorsque plusieurs haplotypes sont à risque. C'est pourquoi, par la suite, nous ne comparerons que les résultats d'EHAP et du TDT_{all}.

[4] un trio est constitué de 2 parents sains et d'un enfant atteint
[5] Pour plus de détails concernant ces deux tests, voir page 15



Modèle	Par.	p-val.
$H_1 : \beta_A, \beta_B, \beta_C, \beta_D, \beta_E, \beta_F, \beta_G, \beta_H, \beta_I, \beta_J, \beta_K$	11	
$H_{0_1} : \beta_E = \beta_C = 0, \beta_A, \beta_B, \beta_D, \beta_F, \beta_G, \beta_H, \beta_I, \beta_J, \beta_K$	10	0,585
$H_{0_2} : \beta_F = \beta_C = 0, \beta_A, \beta_B, \beta_D, \beta_E, \beta_G, \beta_H, \beta_I, \beta_J, \beta_K$	10	0,709
$H_{0_3} : \beta_G = \beta_C = 0, \beta_A, \beta_B, \beta_D, \beta_E, \beta_F, \beta_H, \beta_I, \beta_J, \beta_K$	10	0,636
$H_{0_4} : \beta_H = \beta_A = 0, \beta_G, \beta_B, \beta_D, \beta_E, \beta_F, \beta_C, \beta_I, \beta_J, \beta_K$	10	1,000
$H_{0_5} : \beta_I = \beta_A = 0, \beta_G, \beta_B, \beta_D, \beta_E, \beta_F, \beta_C, \beta_H, \beta_J, \beta_K$	10	0,065
$H_{0_6} : \beta_J = \beta_D = 0, \beta_G, \beta_B, \beta_A, \beta_E, \beta_F, \beta_C, \beta_H, \beta_I, \beta_K$	10	0,499
$H_{0_7} : \beta_K = \beta_B = 0, \beta_G, \beta_A, \beta_D, \beta_E, \beta_F, \beta_C, \beta_H, \beta_J, \beta_I$	10	0,101

Modèle	Par.	p-val.
$H_1 : \beta_1, \beta_2, \beta_3, \beta_4$	4	
$H_{0_1} : \beta_3 = \beta_1 = 0, \beta_2, \beta_4$	3	0,082
$H_{0_2} : \beta_4 = \beta_2 = 0\ , \beta_1, \beta_3$	3	2×10^{-8}

Modèle	Par.	p-val.
$H_1 : \beta_I, \beta_2, \beta_4$	3	
$H_{0_1} : \beta_I = \beta_2 = 0, \beta_4$	2	0,817

FIG. 2.3: Principe de la méthode de Seltman, Devlin et Roeder (EHAP)

Soient 11 haplotypes auxquels correspondent initialement 11 paramètres β.

Au niveau 0 : 7 tests vont être réalisés pour rechercher à la fois si les haplotypes terminaux (E, F, G, H, I, J, K) ont le même risque que les haplotypes internes auxquels ils sont reliés (A, B, C, D) et s'ils sont significativement associés à la maladie. Par exemple, dans le cas de l'haplotype E, on va comparer la vraisemblance du modèle ou $\beta_E = \beta_C$, β_C étant fixé à 0, à la vraisemblance du modèle complet où tous les paramètres sont différents (modèle H_1). Dans notre exemple, aucun test n'est significatif. L'hypothèse nulle d'égalité des paramètres est acceptée pour tous les test.

Au niveau 1 : Après regroupement, il ne reste que 4 paramètres pour le modèle complet H_1, car $\beta_A = \beta_H = \beta_I = \beta_1$, $\beta_B = \beta_K = \beta_2$, $\beta_D = \beta_J = \beta_3$, et $\beta_C = \beta_E = \beta_F = \beta_G = \beta_4$. De la même façon qu'au niveau 0, on réalise, cette fois, 2 tests. Dans notre exemple, le second test est très significatif (marqué d'une étoile sur l'arbre). Le risque associé au groupe 4 est alors significativement différent de celui du groupe 2, qui est lui-même égal à 1. Il y a donc une association entre le gène étudié et la maladie. Supposons de plus que β_4 soit très supérieur à 1, on pourra alors dire que les haplotypes du groupe 4 sont à risque.

Au niveau 2 : Il reste alors 3 paramètres pour le modèle complet (H_1) puisque $\beta_1 = \beta_3 = \beta_I$. Un seul test est réalisé. Il n'est pas significatif, ce qui veut dire que l'hypothèse nulle d'égalité des paramètres β_I et β_2 est acceptée.

En conclusion, la réalisation de 10 tests à un degré de liberté a permis à EHAP d'obtenir des estimations des valeurs des paramètres β associés aux haplotypes. De plus, on a montré l'existence d'une association significative entre le gène étudié et la maladie ($p - value = 2 \times 10^{-8}$). Enfin, l'analyse permet d'identifier le groupe d'haplotypes pour lequel le risque est le plus élevé (groupe 4).

Le TDT_{all} et EHAP ont des puissances comparables qui dépendent essentiellement du modèle simulé. Lorsque tous les haplotypes d'un groupe sont responsables de la susceptibilité à la maladie, EHAP est plus puissant que le TDT_{all}. EHAP est aussi très performant lorsque plusieurs groupes éloignés dans le cladogramme sont à risque. En revanche, lorsque seule une partie des haplotypes d'un groupe est à risque, EHAP est bien moins puissant que le TDT_{all}. Concernant l'effet de la recombinaison, on observe que pour chacun des trois modèles avec recombinaison testés, les résultats de puissance sont très variables. Aucune des deux méthodes ne semble vraiment plus puissante que l'autre. De façon surprenante, on peut noter que EHAP ne semble pas très affecté par les erreurs dans la reconstruction du cladogramme. Cependant, un seul modèle de ce genre est testé, ce qui limite la portée des conclusions que l'on peut en tirer. De façon générale, il faut souligner que les résultats sont extrêmement variables et qu'il y a peu de simulations différentes pour un même type de modèle. De ce fait, il est très difficile d'en tirer des conclusions générales.

Enfin, il faut signaler que certains résultats semblent relativement surprenants. Par exemple, pour le modèle où deux haplotypes (A et H) appartenant au même groupe sont à risque (risque égal pour les deux haplotypes : $e^{\beta_A} = e^{\beta_H} = 1.6$, et risque égal à un pour tous les autres haplotypes), la puissance est de 62.5% pour EHAP et de 71% pour le TDT_{all} alors que pour le modèle où des haplotypes éloignés dans l'arbre (I, A et G) sont à risque ($e^{\beta_I} = e^{\beta_A} = 1.6$ et $e^{\beta_G} = 1.4$, et risque égal à un pour tous les autres haplotypes), qui est *a priori* moins favorable pour une analyse basée sur l'arbre phylogénétique, la puissance de EHAP est de 82% alors que celle du TDT_{all} tombe à 65.5%.

Localisation des sites de susceptibilité à la maladie La méthode EHAP ne permet pas de trouver le ou les marqueurs les plus associés à la maladie. En revanche, elle permet d'identifier le ou les groupes d'haplotypes les plus à risque.

Seltman *et al.* ont étudié si les regroupements d'haplotypes obtenus avec EHAP sont bien cohérents avec les modèles qu'ils ont simulés pour la maladie. Ils observent que pour les modèles simples (un seul groupe responsable de la susceptibilité à la maladie), EHAP parvient à regrouper les haplotypes conformément au modèle simulé dans 60 à 75% des cas. L'identification des haplotypes ou groupes d'haplotypes responsables de la susceptibilité à la maladie est donc acceptable. Dans les cas plus complexes (plusieurs groupes d'haplotypes associés à la maladie, présence de recombinaison, erreurs dans le cladogramme), seulement une partie des haplotypes à risque sont généralement correctement identifiés (35 à 71% des cas).

2.3 La méthode CLADHC (Durrant *et al.*, 2004)

En 2004, Durrant *et al.* ont développé une méthode d'association basée sur l'analyse d'arbres phylogénétiques nommée CLADHC . Elle est très différente de celles que nous avons précédemment décrites : son principe consiste à travailler sur de très longues séquences (plusieurs Mb) en reconstruisant la phylogénie et en calculant une statistique dans des fenêtres de 4 à 10 marqueurs.

2.3.1 Descriptif de la méthode

Comme les méthodes précédentes, CLADHC utilise des données sur des haplotypes de SNPs reconstitués chez des malades et des témoins.

2.3.1.1 Reconstruction phylogénétique

Contrairement à ce que le titre de l'article laisse supposer, Durrant *et al.* (2004) n'utilisent pas une méthode cladistique, mais une méthode basée sur les distances pour reconstruire un arbre. Pour chaque SNP m des haplotypes H_i et H_j (notés $H_i[m]$ et $H_j[m]$), ils définissent une mesure de similitude $s_{i,j}[m]$ de telle sorte qu'elle soit *plus élevée lorsque les allèles partagés sont rares*. Soient A et a les allèles du SNP m, de fréquences respectives f_a et $1 - f_a$. $s_{i,j}[m]$ est définie par :

$$\begin{cases} 1 - f_a & si \quad H_i[m] = H_j[m] = \text{“}a\text{”} \\ f_a & si \quad H_i[m] = H_j[m] = \text{“}A\text{”} \\ f_a(1 - f_a) & si \quad H_i[m] = 0 \quad ou \quad H_j[m] = 0 \\ 0 \quad sinon & \end{cases}$$

CLADHC réalisant l'analyse dans des fenêtres successives de marqueurs, cette mesure de similitude est alors utilisée pour définir une distance entre les haplotypes pour les marqueurs situés dans la fenêtre W :

$$D_{i,j} = 1 - \frac{\sum_{m=1}^{M} s_{i,j}[m] \times w_m}{\sum_{m=1}^{M} w_m}$$

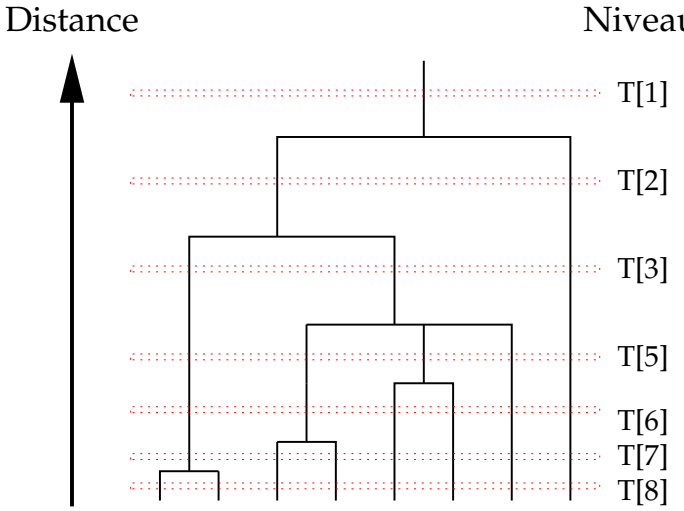
où M est le nombre total de marqueurs formant les haplotypes et w_m est un poids associé à la mesure de similitude $s_{i,j}[m]$ qui vaut 1 si m est situé dans la fenêtre W et 0 sinon. Les valeurs de w_m permettent de considérer indépendamment chaque fenêtre comme étant formée d'un ensemble de SNPs entre lesquels il n'y a pas de recombinaison.

À partir de ces distances, un arbre est reconstruit par une méthode de classification hiérarchique combinatoire (type UPGMA), en regroupant à chaque étape les haplotypes ou les groupes d'haplotypes séparés par la plus faible distance. Pour ce type de méthodes de reconstruction, le problème de l'enracinement de l'arbre ne se pose donc pas car l'arbre obtenu s'enracine par construction.

2.3.1.2 Analyse de l'arbre

Principe L'arbre obtenu est découpé en plusieurs niveaux ($T[1]$ au niveau de la racine, jusqu'à $T[c]$ au niveau des feuilles), chacun formé de c groupes ou *clusters* d'haplotypes (voir Figure 2.4).

L'analyse des données sur l'arbre obtenu dans une fenêtre W est réalisée en utilisant un modèle de régression logistique faisant intervenir le risque de chaque haplotype en fonction du *cluster* auquel il appartient et éventuellement, de l'effet de co-variables (sexe, effets environnementaux, etc.). Pour chaque niveau, la vraisemblance est calculée et un test de rapport de vraisemblance est réalisé en comparant la vraisemblance au niveau $T[c]$ à la vraisemblance au niveau $T[1]$, le niveau $T[1]$ correspondant à l'hypothèse nulle où tous les haplotypes appartiennent à un seul *cluster* et ont donc les mêmes risques d'être portés par des



Les pointillés rouges recoupent les c clusters d'haplotypes existant au niveau $T[c]$.

FIG. 2.4: Découpage de l'arbre dans la méthode de Durrant *et al.* (CLADHC)

malades ou des témoins. Pour chaque fenêtre, le niveau le plus significatif (i.e. celui où la p-value du test de rapport de vraisemblance est la plus faible) est retenu et appelé $T[MAX]$.

Cas des haplotypes rares Dans les niveaux les plus élevés de l'arbre, un problème de petits effectifs peut se poser. Durrant *et al.* (2004) ont proposé de rassembler à chaque niveau les *clusters* totalisant moins de 5% des individus (malades + témoins). Il est alors possible d'estimer un risque associé à ce groupement de *clusters* et de le traiter dans l'analyse comme tous les autres *clusters*.

Correction pour les tests multiples Dans l'analyse proposée par Durrant *et al.*, il existe deux niveaux de tests multiples :

- dans une fenêtre W, le test de rapport de vraisemblance est réalisé pour tous les niveaux de l'arbre ;
- sur l'ensemble des séquences, N fenêtres chevauchantes sont analysées.

Le programme développé par Durrant *et al.* (2004) permet soit d'évaluer des p-values corrigées par permutations, soit d'utiliser la correction de Bonferroni pour les tests multiples.

2.3.2 Étude de puissance

Durrant *et al.* (2004) ont étudié la puissance de leur méthode en réalisant des simulations sous différents modèles génétiques et pour différentes tailles de fenêtres. Ils ont comparé les résultats obtenus avec leur méthode à deux autres méthodes classiques :

- un test allélique réalisé pour chaque locus par un χ^2 de Pearson, corrigé par le nombre de SNPs analysés ;
- un test haplotypique basé sur le même modèle de régression logistique que le test phylogénétique. Ce test correspond en fait à l'analyse du niveau le plus élevé de l'arbre,

où tous les clusters ne contiennent qu'un seul haplotype. Ce test est corrigé pour le nombre de fenêtres analysées.

La puissance est calculée sur les fenêtres situées dans la région flanquante du gène de susceptibilité à la maladie tandis que l'erreur de type I est évaluée sur les autres fenêtres.

2.3.2.1 Simulation des données

Pour générer leurs données, Durrant *et al.* ont utilisé des données de familles du CEPH génotypées sur une région de 10 Mb située sur le chromosome 20. Ces données leur ont fourni 92 haplotypes qu'ils ont utilisés comme base pour fabriquer de nouveaux haplotypes pour 1000 malades et de 1000 témoins. D'après Durrant *et al.* (2004), ce processus leur permet d'obtenir de nouveaux haplotypes dont le déséquilibre de liaison est réaliste. Il est important de noter que, dans le processus de simulation, le locus à risque est éliminé des nouveaux haplotypes à la dernière étape de leur fabrication.

Le processus de simulation est décrit sur la Figure 2.5.

La fabrication des nouveaux haplotypes débute par le choix du SNP de susceptibilité au hasard parmi les 5216 marqueurs disponibles. Le SNP est choisi en fonction de la fréquence souhaitée pour l'allèle de susceptibilité à la maladie.
Les chromosomes des 1000 malades et des 1000 témoins vont être alors générés. Prenons l'exemple de la fabrication des chromosomes d'un individu témoin. Le génotype de cet individu au locus de susceptibilité (nommé L) est tout d'abord choisi en fonction des valeurs des pénétrances de la maladie. Chaque allèle est ensuite attribué aléatoirement à un chromosome. Enfin, tous les locus de chaque chromosome sont remplis selon le processus suivant :

1. *Simulation des 4 SNPs autour du locus de susceptibilité L :* l'un des 92 haplotypes du CEPH est tiré aléatoirement. Si l'allèle au site de susceptibilité L de cet haplotype est le même que celui de l'haplotype en cours de construction, alors les allèles des sites $L-2$, $L-1$, $L+1$ et $L+2$ de l'haplotype du CEPH seront affectés aux mêmes locus de l'haplotype en cours de construction. Sinon, recommencer cette étape.
2. *Simulation du SNP $L+3$:* un autre des 92 haplotypes du CEPH est tiré aléatoirement. Si les allèles de cet haplotype sont les mêmes que ceux de l'haplotype en construction pour les SNPs $L-1$ à $L+2$, alors l'allèle du site $L+3$ de l'haplotype du CEPH est affecté au locus $L+3$ de l'haplotype en construction. Sinon, recommencer cette étape.
3. Recommencer l'étape 2 pour tous les ensembles de 5 SNPs chevauchants $[L, L+4]$ (remplissage du locus $L+4$), $[L-3, L+1]$ (remplissage du locus $L-3$), ..., jusqu'à avoir rempli tous les SNPs de l'haplotype en construction.
4. Éliminer le SNP de susceptibilité de l'haplotype.

FIG. 2.5: Processus de simulation des haplotypes par Durrant *et al.* (2004)

2.3.2.2 Résultats

Durrant *et al.* (2004) ont réalisé des simulations pour différentes tailles de fenêtre (entre 4 et 10 marqueurs) et pour différents modèles génétiques (fréquences de l'allèle de susceptibilité entre 0.01 et 0.5, risque relatif des homozygotes (GRR_{hom}) entre 1 et 40, risque relatif des hétérozygotes (GRR_{het}) entre 1 et 3.

L'erreur de type I Par simulations, Durrant *et al.* (2004) ont montré que les trois tests étaient conservatifs avec des valeurs de l'erreur de type I inférieures à 5% pour une valeur nominale fixée à 5%. Du fait du double niveau de correction de Bonferroni, le test phylogénétique est le plus conservateur (erreur de type I toujours inférieure à 1,75%). Les simulations ont montré que le modèle maladie a peu d'influence sur l'erreur de type I, ce qui est attendu car l'erreur de type I est évaluée sur une région ne contenant pas le gène de susceptibilité. La taille de la fenêtre a une influence sur le test haplotypique et sur le test phylogénétique : plus les fenêtres sont constituées d'un grand nombre de marqueurs, plus les tests sont conservateurs. Ce résultat s'explique très simplement : dans des fenêtres contenant de nombreux marqueurs, le nombre d'haplotypes différents sera plus élevé, ce qui va augmenter le nombre de degrés de liberté du test haplotypique et le nombre de niveaux dans l'arbre (i.e. le nombre de tests réalisés) pour le test phylogénétique.

La puissance L'étude de puissance a été réalisée pour une fréquence de l'allèle de susceptibilité égale à 0,05. Lorsque les risques relatifs génotypiques sont faibles ou modérés ($GRR_{het} < 2$ et $GRR_{hom} < 10$), Durrant *et al.* (2004) observent que la puissance des trois tests est équivalente et quasi nulle. Lorsque les risques relatifs augmentent, ils observent que la puissance augmente pour les trois tests. Elle reste cependant faible, 20% au maximum pour $GRR_{het} = 3$ et $GRR_{hom} = 40$, pour le test allélique alors qu'elle atteint environ 97% pour le test phylogénétique et 90% pour le test haplotypique. Quel que soit le modèle génétique, ils observent que le test phylogénétique est toujours le plus puissant.

Durrant *et al.* (2004) ont aussi étudié l'influence de la taille de la fenêtre sur la puissance de la méthode phylogénétique. Ils ont montré qu'il existait un optimum pour la taille de la fenêtre correspondant à 6 marqueurs pour leur jeu de données. Cependant, cette valeur varie selon les jeux de données, en particulier en fonction du déséquilibre de liaison entre les différents marqueurs.

2.3.3 Localisation du site de susceptibilité

Durrant *et al.* (2004) ont montré qu'en plus de détecter une association entre un gène et une maladie, leur méthode permet de localiser la région contenant le site de susceptibilité, même si ce dernier n'est pas génotypé. Pour chaque fenêtre analysée sur le gène, `CLADHC` donne une p-value. En représentant graphiquement l'inverse du logarithme en base 10 de la p-value de chaque fenêtre en fonction de leur localisation sur le chromosome, on peut facilement identifier un pic correspondant à la plus petite p-value et donc à la région dans laquelle se trouve probablement le locus de susceptibilité.

Durrant *et al.* (2004) ont appliqué cette méthode sur un jeu de données concernant la mucoviscidose et le gène CFTR (*Cystic Fibrosis Transmembrane conductance Regulator*). Ces données (Kerem *et al.*, 1989) comportent 94 haplotypes de malades et 92 haplotypes de témoins constitués de 23 marqueurs bialléliques, les plus distants étant séparés par 1,8Mb. Ces données ont été analysées à l'aide de la méthode `CLADHC` en utilisant une taille de fenêtre fixée à 6 marqueurs. Ils ont trouvé un pic où la p-value vaut 10^{-20} à la position 879 kb, qui correspond, à 1kb près, à la position de la mutation $\Delta F508$, connue pour être responsable de la mucoviscidose.

2.4 La méthode de Tree Scanning (Templeton *et al.*, 2005)

Toutes les méthodes précédemment développées permettent d'analyser des données haplotypiques mais ne permettent pas de prendre en compte d'éventuels effets du génotype, en particulier des interactions entre les haplotypes d'un individu sur le phénotype. De plus, il est relativement difficile d'attribuer avec précision les effets de chaque haplotype d'un individu. Afin de pallier ce problème, Templeton *et al.* (2005) ont développé la méthode de Tree Scanning. Elle a été implémentée dans le logiciel `TreeScan` (Posada *et al.*, 2005).

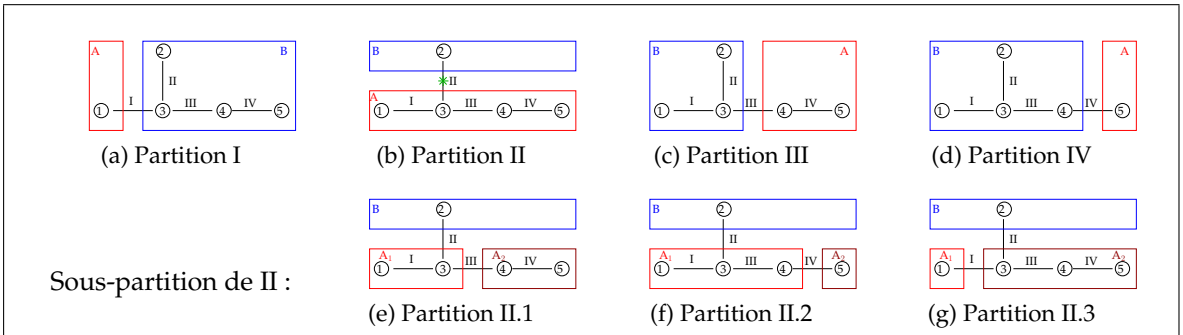
2.4.1 Description de la méthode

Elle permet de réaliser des études d'association sur des données quantitatives en population. Comme les méthodes décrites précédemment, elle utilise des haplotypes de SNPs.

2.4.1.1 Reconstruction phylogénétique

La reconstruction phylogénétique est réalisée avec le logiciel `TCS` (Clement *et al.*, 2000) qui implémente l'algorithme de parcimonie développé par Templeton *et al.* (1992) (voir page 37).

2.4.1.2 Analyse de l'arbre



Premier niveau de partition : pour chaque partition, les haplotypes situés dans la boîte rouge seront recodés en A, alors que les haplotypes de la boîte bleue seront recodés en B. Par exemple, dans le cas de la partition III, un individu portant les haplotypes 1 et 5 sera affecté à la classe AB.

Deuxième niveau de partition : une association significative a été identifiée pour la partition II. Le groupe A est donc, à son tour, partitionné en deux sous-groupes A_1 et A_2.

FIG. 2.6: Principe de la méthode de Tree Scanning

Principe Chaque branche de l'arbre permet de définir une partition de l'arbre, c'est-à-dire deux groupes d'haplotypes : le groupe A et le groupe B (voir Figure 2.6). Pour chaque partition, connaissant les haplotypes de chaque individu, il est possible de recoder le génotype d'un individu en fonction du groupe auquel appartiennent ses haplotypes. On obtient donc trois classes d'individus (AA, AB et BB) dans lesquelles sont réparties les valeurs du trait

quantitatif associé aux individus. Il est alors possible de tester l'association entre le trait et la partition en réalisant une analyse de variance classique. Dans tous les cas, la p-value associée au test est calculée par permutation entre les valeurs du trait dans les trois classes considérées.

Par la suite, lorsqu'une association est détectée au niveau d'une partition, les deux groupes définis sur l'arbre sont à leur tour successivement partitionnés. Par exemple, A peut être partitionné en A_1 et A_2 (voir Figure 2.6). L'association génotype/phénotype est alors testée de la même manière que précédemment, mais sur un système à 6 groupes[6] au lieu de trois. Le découpage de l'arbre se poursuit jusqu'à ce que toutes les partitions et sous-partitions significatives aient été détectées.

Lorsque plusieurs associations sont détectées au niveau de branches contiguës dans l'arbre, il est probable qu'une seule partition a un effet qui "déborde" sur les branches adjacentes. Du coup, seule la branche pour laquelle la p-value est la plus faible est considérée comme significative et conservée pour l'analyse des niveaux supérieurs.

Cas des haplotypes rares Lorsque les partitions du type "un haplotype versus tous les autres" sont réalisées, le problème des petits effectifs peut se poser. Dans le programme `TreeScan` (Posada *et al.*, 2005), il est possible de définir un seuil d'effectifs en dessous duquel le test n'est pas réalisé.

Correction pour les tests multiples La méthode de Tree Scanning réalise un grand nombre de tests (un pour chaque partition de l'arbre, pour chaque niveau de partition). Templeton *et al.* proposent alors de réaliser une correction pour chaque niveau de partition basée sur la procédure de permutation de Westfall et Young (1993) (voir page 19).

2.4.2 Étude de puissance

La puissance de cette méthode n'a jamais été testée par simulations. Cependant, son efficacité a été comparée à celle de l'analyse de clades emboîtés sur un jeu de données concernant l'activité ADH (Alcool deshydrogénase) chez la drosophile. Templeton *et al.* (2005) ont montré que les deux méthodes donnent des résultats comparables. Templeton *et al.* (2005) l'ont aussi comparée à une analyse génotypique locus par locus sur un jeu de données concernant les taux d'ApoE, de triglycérides, de cholestérol et d'HDL-C et le gène APOE. Pour cela, ils ont comparé les SNPs détectés comme étant significativement associés avec les traits étudiés par la méthode locus par locus, avec les SNPs qui mutent sur les branches détectées par `TreeScan` . Il ont montré que les deux méthodes sont largement concordantes, mais que la méthode de `TreeScan` permet de détecter plus de SNPs associés que la méthode génotypique. Templeton *et al.* (2005) en concluent donc que l'utilisation de la phylogénie permet d'augmenter la puissance de détection d'associations. Cependant, comme ils utilisent des données réelles pour lesquelles les effets ne sont pas encore connus avec certitude, il semble difficile de dire si les associations nouvelles détectées par la méthode de `TreeScan` sont réelles ou si elles correspondent à des faux positifs.

[6] Avec A_1, A_2 et B, les 6 groupes possibles sont alors A_1A_1, A_1A_2, A_1B, A_2A_2, A_2B et BB

2.5 La méthode de Tzeng (Tzeng, 2005)

Dans une première étude, Tzeng *et al.* (2003) ont développé une méthode de regroupement des catégories rares consistant à rassembler chaque haplotype rare avec l'haplotype fréquent dont il est séparé par un seul pas évolutif. Si un tel haplotype fréquent n'existe pas, l'haplotype rare est alors regroupé avec tous les autres haplotypes rares qu'il n'a pas été possible de classer. Cette méthode comporte quelques faiblesses :

- le critère pour déterminer si un haplotype est rare ou fréquent est un seuil de fréquence, déterminée par simulation sur un jeu de données précis ;
- la catégorie regroupant les haplotypes rares non séparés par un seul pas évolutif d'haplotypes fréquents est très hétérogène.

Tzeng (2005) a alors développé une nouvelle méthode qui utilise davantage l'histoire évolutive des haplotypes du jeu de données.

2.5.1 Descriptif de la méthode

La méthode proposée par Tzeng (2005) est conçue pour faire des études d'association cas/témoins sur des haplotypes. Elle est basée sur l'identification d'un ensemble d'haplotypes fréquents, $H*$, qui vont constituer la base sur laquelle les haplotypes plus rares vont se greffer. Les haplotypes de la base $H*$ n'étant pas reliés entre eux, il n'y a pas vraiment de reconstruction d'un arbre phylogénétique ou d'un réseau pour cette méthode.

Détermination de la base des haplotypes fréquents, $H*$ Cette base est constituée des l haplotypes les plus fréquents. Il faut alors déterminer le nombre optimal d'haplotypes l de telle sorte qu'il y en ait suffisamment pour optimiser la quantité d'informations, mais pas trop afin de limiter l'augmentation du nombre de degrés de liberté du test et donc, la perte de puissance. Pour cela, Tzeng utilise un critère basé sur l'information nette[7].

Rattachement des haplotypes restants à la base $H*$ Les haplotypes sont tout d'abord partitionnés en différentes catégories $H^{(1)}$, $H^{(2)}$, ..., $H^{(j)}$ en fonction du nombre de pas qui les séparent d'un des haplotypes de la base $H*$. Puis, en commençant avec $j = j_{max}$, chaque élément du groupe $H^{(j)}$ est regroupé avec un haplotype séparé par un seul pas évolutif appartenant au groupe $H^{(j-1)}$. Par étapes successives, les haplotypes vont donc finalement être regroupés avec un des haplotypes de la base $H*$. Lorsqu'un haplotype de $H^{(j)}$ peut être groupé avec plusieurs haplotypes de $H^{(j-1)}$, des probabilités de regroupement avec chaque haplotype concerné sont alors assignées. Ces probabilités sont basées sur la fréquence des haplotypes de $H^{(j-1)}$: plus un haplotype est fréquent, plus il est supposé ancien, et plus la probabilité de regroupement avec cet haplotype est élevée.

Test d'association Une fois tous les haplotypes rares regroupés avec l'un des haplotypes de $H*$, on dispose de l catégories haplotypiques. L'association entre le gène et la maladie est testée en réalisant un test de χ^2 de Pearson à $l-1$ degré de liberté de la même manière que pour un test d'association haplotypique classique (voir page 15).

[7] Pour un ensemble d'haplotypes h_i de fréquences f_{h_i} ,l'information nette est obtenue en calculant l'entropie ($-\sum_{i=1}^{l} f_{h_i} \times log_2 f_{h_i}$) et en lui soustrayant une pénalité liée à l'augmentation du nombre de catégories

2.5.2 Étude de puissance

La puissance du test d'association de Tzeng (2005) (notée H_{groupe} par la suite) a été estimée par simulations et comparée à deux autres tests d'association :

- un test d'association haplotypique classique prenant en compte tous les haplotypes : H_{total} ;
- un test d'association où seuls les haplotypes constituant la base $H*$ ont été conservés. Les autres haplotypes ont été éliminés de l'analyse : $H_{tronque}$.

2.5.2.1 Méthode de simulation

Les haplotypes ont été simulés en utilisant un modèle de coalescence avec recombinaison (programme MS de Hudson (2002)). Les paramètres ont été choisis de manière à ce que les déséquilibres de liaison ressemblent à ceux observés pour les gènes *IL10RB* et *SELP* dans la base de données SeattleSNPs. Au total, 100 séquences de SNPs ont été générées pour chaque gène. Parmi les SNPs constituant ces séquences, un site de susceptibilité est choisi en fonction de sa fréquence allélique (0,1 ; 0,3 ou 0,5) et de sa position sur le gène (dans un bloc haplotypique où il y a peu de recombinaisons ou bien au niveau d'un point chaud de recombinaisons). Seuls les 2 marqueurs précédant et les 3 marqueurs suivant le site de susceptibilité sont conservés pour constituer des haplotypes formés de 6 SNPs (dont le site de susceptibilité). Enfin, deux séries de N individus (N malades et N témoins, $N = 100 ou 200$) sont générées en tirant aléatoirement avec remise deux haplotypes parmi les 100 possibles et en attribuant un statut maladie à ce génotype en fonction des risques relatifs (f_1=2 pour les hétérozygotes et f_2=3 pour les homozygotes porteurs) et de la prévalence[8] (K=0.01) de la maladie.

2.5.2.2 Résultats

L'erreur de type I Pour les trois tests, les erreurs de type I sont globalement comprises entre 2,8% et 6,6%. Les différences entre les tests peuvent être très importantes, en particulier dans le cas où le site de susceptibilité est situé dans un point chaud de recombinaison. Globalement, le test H_{groupe} possède l'erreur de type I la plus élevée et le test H_{total} l'erreur de type I la plus faible. Du fait des différences importantes entre les erreurs de type I, il sera parfois difficile de comparer les puissances obtenues.

La puissance La comparaison de la puissance des trois tests révèle que le test H_{total} est moins puissant que les deux autres tests, sans doute à cause de son plus grand nombre de degrés de liberté. Le test H_{groupe} est quasiment toujours légèrement plus puissant que le test $H_{tronque}$, mais la différence n'est significative à 5% que dans 4 conditions sur les 36 testées, et à 1% que dans une des 36 conditions testées. De plus, comme l'erreur de type I est généralement plus forte pour H_{groupe}, il est difficile de dire si le regroupement des haplotypes permet un gain de puissance significatif par rapport à l'élimination des haplotypes rares.

Concernant les variations de puissance en fonction des paramètres des simulations, on peut noter que des résultats comparables sont obtenus pour les deux "modèles" testés (IL10RB et SELP). Comme attendu, on observe que la puissance augmente avec la taille de

[8] La prévalence correspond à la fréquence de la maladie dans une population donnée

l'échantillon. On peut aussi noter que la puissance des trois tests augmente avec la fréquence de l'allèle de susceptibilité et lorsque le site de susceptibilité est dans un bloc haplotypique.

2.5.2.3 Analyse de données

Tzeng a aussi montré la puissance de son test par rapport aux tests H_{total} et $H_{tronque}$ en analysant 8 SNPs intragéniques du jeu de données de Hugot *et al.* (2001) sur le gène CARD15 et la maladie de Crohn (pour une description plus complète de ce jeu de données, voir page 128). Il a évalué les p-values associées aux trois tests pour ces données et montré que la p-value la plus faible était obtenue pour le H_{groupe} ($p = 1,41 \times 10^{-10}$). Le test $H_{tronque}$ a une p-value 5 fois plus élevée ($p = 6,67 \times 10^{-10}$) pour le même nombre de degrés de liberté (12) et le test H_{total} est toujours moins puissant ($p = 3.79 \times 10^{-7}$).

2.6 Utilisation des méthodes phylogénétiques sur des données

Depuis les travaux de Templeton *et al.* (1987), plusieurs jeux de données ont été analysés avec des méthodes basées sur les phylogénies. Cependant, jusqu'en 2003, aucun logiciel permettant de réaliser ces analyses n'était distribué. Les différents auteurs de ces analyses ont donc adapté le principe de ces méthodes à leurs données. Dans cette partie, nous allons récapituler les principales caractéristiques des analyses réalisées et nous allons pointer les différences méthodologiques majeures entre les différentes études sur trois points principaux : la méthode de reconstruction de l'arbre, le type d'analyse statistique réalisé et l'utilisation éventuelle de la méthode pour localiser le site de susceptibilité.

La liste des analyses et leurs caractéristiques est exposée dans deux Tableaux : Tableau 2.1 pour les analyses de données quantitatives et Tableau 2.2 pour les analyses de données cas/témoins.

2.6.1 Reconstruction de l'arbre

Dans la plupart des études, des méthodes de parcimonie ont été utilisées pour reconstruire la phylogénie des haplotypes. Certains auteurs ont utilisé des logiciels classiques de reconstruction d'arbres phylogénétiques tels que `PAUP` et `PHYLIP` (Hallman *et al.*, 1994; Kittles *et al.*, 1999; Darlu et Génin, 2001; Templeton *et al.*, 1987; Zhu *et al.*, 2000, 2001) ou de réseaux tels que les logiciels `NETWORK` Bandelt *et al.* (1995) (utilisé par Ferjerman *et al.* (2004)) ou `TCS` Clement *et al.* (2000) (utilisé par (Templeton *et al.*, 2005)). Dans certaines études, du fait des faibles nombres de marqueurs et d'haplotypes étudiés, les arbres sont reconstruits "à la main", en reliant les haplotypes séparés par un seul pas évolutif. Lorsque des haplotypes sont séparés par plus d'un pas évolutif, la plupart des auteurs ont choisi d'ajouter un ou plusieurs haplotypes non observés dans l'échantillon afin d'obtenir un unique arbre (Friedlander *et al.*, 1995; Haviland *et al.*, 1995; Templeton *et al.*, 1988; Wang *et al.*, 2003). Keavney *et al.* (1998) ont, pour leur part, choisi de ne pas relier les 3 groupes d'haplotypes qu'ils avaient obtenus. Enfin, Templeton *et al.* (1992); Crandall et Templeton (1993) ont développé des règles pour construire le/les cladogrammes les plus probables et pour résoudre les boucles dues à l'homoplasie (voir page 37). Quelques études utilisent ces règles pour reconstruire la phylogénie des haplotypes (Lobos et Todd, 1997, 1998; Templeton et Sing, 1993; Templeton, 1995).

Les arbres reconstruits et analysés dans les différentes études sont toujours non enracinés. Seuls Darlu et Génin (2001) ont travaillé à la fois sur des arbres enracinés et non enracinés. Ils ont montré que les deux méthodes donnaient des résultats équivalents.

2.6.2 Type d'analyse statistique réalisée

Selon le type de données, différents tests statistiques sont possibles. Pour les données quantitatives, des analyses de variances (ANOVA ou NANOVA) ou des régressions linéaires sont réalisées. Pour les données qualitatives, des tests de χ^2 ou des régressions logistiques sont utilisés.

Certaines études sont des applications directes des méthodes développées précédemment dans ce chapitre (Durrant *et al.*, 2004; Templeton *et al.*, 2005). Pour les autres, la base de l'analyse statistique de l'arbre est très clairement la méthode d'analyse des clades emboîtés de Templeton. Certains articles sont très fidèles à cette méthode et la suivent scrupuleusement (Lobos et Todd, 1997, 1998; Kittles *et al.*, 1999). Mais la plupart des auteurs ont adapté la méthode au traitement de leurs données. Les auteurs qui utilisent des modèles de régression linéaire (Hallman *et al.*, 1994; Keavney *et al.*, 1998; Wang *et al.*, 2003; Zhu *et al.*, 2000, 2001) utilisent une méthode du type de celle développée par Seltman *et al.* (2003). Certains auteurs ne font pas une analyse emboîtée mais utilisent la structure de l'arbre pour définir les groupes qu'ils vont comparer (Darlu et Génin, 2001; Haviland *et al.*, 1995). C'est aussi le cas d'une des analyses de Ferjerman *et al.* (2004) : ils ont utilisé un réseau médian réduit (*reduced median network*) pour définir 7 haplogroupes pour lesquels ils ont calculé la valeur du taux d'angiotensinogène (AGT). Les taux d'AGT de chaque individu sont ensuite permutés 1000 fois afin d'évaluer la significativité des taux d'AGT associés aux sept haplogroupes. Ils ont aussi utilisé l'arbre obtenu pour tester la corrélation entre la distance génétique entre deux haplotypes et le taux d'AGT. Enfin, Katsov *et al.* (2004) ont utilisé une méthode permettant de prendre en compte l'effet des génotypes des individus. Cette méthode est proche de celle décrite par Templeton *et al.* (2005), mais l'arbre qu'ils étudient ne comportant que trois groupes, ils n'analysent qu'une seule partition de l'arbre : celle dans laquelle les trois groupes sont séparés. Ensuite, comme dans la méthode de Templeton *et al.* (2005), le génotype de chaque individu est recodé pour donner un "cladetype"[9] en fonction des groupes auxquels appartiennent ses 2 haplotypes. Pour les traits quantitatifs, les moyennes des valeurs des traits sont comparées par une ANOVA entre les différents "cladetypes". Pour les traits qualitatifs, les risques associés à chaque "cladetype" sont évalués par régression logistique. Grâce à cette méthode, ils ont mis en évidence des associations entre le gène *ACE* et différents traits quantitatifs (indice de masse corporelle : $\frac{Poids}{Taille^2}$, taux de glucose plasmatique, niveau d'insuline, etc.).

2.6.3 Détection de l'association

Dans toutes ces études, l'analyse phylogénétique a généralement permis de mettre en évidence une ou plusieurs associations entre des haplotypes ou des groupes d'haplotypes et le trait étudié. Quelques études ont comparé les résultats obtenus par une analyse basée sur la phylogénie avec ceux obtenus par des analyses classiques. Haviland *et al.* (1995) ont

[9] Comme fréquemment dans la littérature, le mot "clade" est employé même lorsque l'arbre n'est pas enraciné, ni même reconstruit par une méthode cladistique

montré que, sur leurs données, une analyse de variance globale sur l'ensemble des haplotypes n'est pas significative alors que les contrastes définis par la phylogénie des haplotypes permettent de détecter une association. En revanche, Rodriguez *et al.* (2004) ont montré que des méthodes haplotypiques classiques donnent des p-value dix fois plus élevées que EHAP .

2.6.4 Localisation des sites de susceptibilité

Certaines études ont aussi utilisé l'arbre phylogénétique pour localiser plus précisément le/les marqueurs impliqués dans le déterminisme du trait. Parmi elles, deux méthodes sont utilisées[10].

La première consiste à étudier les haplotypes recombinants comme le propose Templeton (Templeton *et al.*, 1987; Templeton et Sing, 1993) (voir page 38). En utilisant cette méthode sur des données provenant d'une population européenne, Keavney *et al.* (1998) ont montré qu'un variant impliqué dans le déterminisme du taux d'ACE (Angiotensin-I Converting Enzyme) se situait entre les marqueurs T-93C et T1237C (en aval d'un point de recombinaison situé entre ces marqueurs). Zhu *et al.* (2000) ont approfondi cette étude en travaillant sur 8 nouveaux polymorphismes dans une population africaine. Ils ont trouvé deux points de recombinaison qui leur ont permis de limiter la région contenant le variant fonctionnel à une région située entre les marqueurs 20833 et 22982. L'avantage de cette méthode est qu'elle ne demande pas que le variant fonctionnel soit typé. Par contre, elle nécessite la présence de recombinaisons, ce qui entraîne des difficultés pour la reconstruction de l'arbre, ainsi que l'identification fiable et précise des haplotypes recombinants et des points de recombinaison.

La seconde méthode de localisation consiste à s'intéresser aux marqueurs qui mutent le long des branches définissant les groupes associés au trait étudié. En utilisant cette méthode, Darlu et Génin (2001) ont identifié un des sites de susceptibilité simulé pour les données du GAW 12 (Genetic Analysis Workshop) parmi l'ensemble des sites définissant un groupe contenant un excès d'haplotypes portés par des malades. Ferjerman *et al.* (2004) ont aussi montré l'intérêt de cette méthode en comparant les résultats d'une analyse locus par locus à l'analyse haplotypique. Ils ont observé que les marqueurs les plus fortement associés aux taux d'AGT sont ceux qui définissent les haplogroupes associés à des taux extrêmes d'AGT. Cette méthode ne fait pas du tout intervenir la recombinaison et peut donc être mise en œuvre dans des régions très courtes où il n'y a pas de recombinaisons. En revanche, cette méthode implique que le variant fonctionnel soit génotypé.

[10] on ne parlera pas ici de la méthode de Durrant *et al.* (2004), déjà décrite au paragraphe 2.3.3, page 48 car, jusqu'à présent, elle n'a pas été utilisée sur d'autres données que celles utilisées par Durrant *et al.* (2004)

Article	Région	Trait étudié	marqueurs	Nb haplo	Nb ind	Reconstr. arbre	Méthode analyse
Templeton *et al.* (1987, 1988, 2005)	*ADH* Drosophile	taux ADH	?	25	49 chromosomes	`PHYLIP` `TCS`	NANOVA permutations Tree Scan
Templeton *et al.* (1988)	*APO-AI-CIII-AIV*	taux TG et cholestérol	3	5	88	à la main	contrastes non emboîtés
Templeton et Sing (1993)	α-*amylase* Drosophile	activité α-amylase	20	21	44	algorithme Templeton	NANOVA
Hallman *et al.* (1994)	*ApoB*	taux ApoB et lipides plasmatiques	5	11	121 familles	`PAUP`	modèle linéaire
Haviland *et al.* (1995)	*APO-AI-CIII-AIV*	taux lipides, lipo-protéines plasmatiques	6	7	140	à la main	contrastes non emboîtés
Keavney *et al.* (1998)	ACE	taux de ACE	10	7 plus fréquents	555 membres de 83 familles	à la main	modèle linéaire
Kittles *et al.* (1999)	chr. Y non recomb.	traits associés à alcoolisme	8	102	132 cas 227 témoinss	`PAUP`	NANOVA
Zhu *et al.* (2000)	ACE	taux de ACE	7	8 plus fréquents	15 Afro-Carribean 98 Européens	`PHYLIP`	modèle linéaire
Zhu *et al.* (2001)	données simulées pour GAW12		14	10 plus fréquents	495	`PHYLIP`	modèle linéaire
Wang *et al.* (2003)	*APOA4*	taux lipides plasmatique	5	4 ou 6 selon pop	400 à 700 selon pop	à la main	modèle linéaire

Article	Région	Trait étudié	marqueurs	Nb haplo	Nb ind	Reconstr. arbre	Méthode analyse
Rodriguez *et al.* (2004)	*IGF2-INS-TH*	BMI, TG, pression artérielle, etc.	3 (dont 1 à 5 allèles)	10 plus fréquents	2743 hommes sains	EHAP	EHAP
Ferjerman *et al.* (2004)	AGT	taux AGT plasmatique	19	104 (7 haplogroupes)	57	réseau médian réduit Bandelt *et al.* (1995)	contrastes non emboîtés
Katsov *et al.* (2004)	*ACE*	20 traits (BMI, taux *APOB*, LDL-C, etc.)	2	3 groupes	> 2000	littérature (Keavney *et al.*, 1998)	ANOVA
Templeton *et al.* (2005)	*ApoE*	taux ApoE cholestérol, etc.	21	39	330 à 840 selon pop	TCS Clement *et al.* (2000)	Tree scan (ANOVA)

ACE : Angiotensin-I Converting Enzyme
ADH : Alcohol Dehydrogenase
AGT : Angiotensinogen
APO- : Apolipoprotein-
BMI : Body Mass Index
DRD2 : Dopamin Receptor D2
LDL-C : Low Density Lipoprotein C
MMD : maladies maniaco-dépressives
(N)ANOVA : (Nested) Anlysis of Variance
TG : triglycérides
TH : Tyrosine Hydroxylase

TAB. 2.1: Revue des analyses phylogénétiques sur des données quantitatives

Article	Région	Trait étudié	marqueurs	Nb haplo	Nb ind	Reconstr. arbre	Méthode analyse
Templeton (1995)	*ApoE*	maladie d'Alzheimer	5	15	36 cas/38 témoins 34 cas/ 36 témoins	algorithme Templeton	χ^2 emboîtés
Lobos et Todd (1997)	*TH*	MMD alcoolisme	4	9	81 MMD[a] 113 alcooliques[a] 80 témoins[a]	algorithme Templeton	χ^2 emboîtés
Lobos et Todd (1998)	*DRD2*	alcoolisme	5	6	55 cas 80 témoins	algorithme Templeton	χ^2 emboîtés
Kittles *et al.* (1999)	chr. Y non recomb.	alcoolisme	8	102	132 cas 227 témoins	PAUP	χ^2 emboîtés
Darlu et Génin (2001)	données simulées pour GAW12		281 64	178 48	G1 : 33 cas 56 témoins G2 : 38 cas 68 témoins	PAUP	χ^2 entre groupes majeurs
Katsov *et al.* (2004)	ACE	Alzheimer infarctus du myocarde	2	3 groupes	> 2000	littérature (Keavney *et al.*)	régression logistique
Durrant *et al.* (2004)	*CFTR*	mucoviscidose	23	22	94 cas 92 témoins	méthode de distance	CLADHC (reg logistique)
Strauss *et al.* (2005)	*BNDF*	troubles de l'humeur chez l'enfant	2	4	258 familles	EHAP	EHAP

[a] : mélange de populations blanches non hispaniques d'origine ethnique variée

TH : Tyrosine Hydroxylase — MMD : maladies maniaco-dépressives

DRD2 : Dopamine Récepteur D2 — *APO-* : Apolipoprotein-

ACE : Angiotensin-I Converting Enzyme — G1/G2 : gène 1, gène 2

CFTR : Cystic Fibrosis Transmembrane conductance Regulator

BNDF : Brain Derived Neurotrophic Factor —

TAB. 2.2: Revue des analyses phylogénétiques sur des données qualitatives

2.7 Bilan sur les méthodes d'association et de localisation de sites de susceptibilité basées sur des phylogénies

Les méthodes décrites précédemment utilisent toutes un arbre des haplotypes pour réaliser des tests d'association et éventuellement pour localiser les variants responsables de la susceptibilité à la maladie sur le gène. Cependant, les méthodes sont relativement différentes. Pour la reconstruction de l'arbre, Templeton *et al.* (1987, 2005) utilisent leur propre algorithme, tandis que Seltman *et al.* (2003) et Durrant *et al.* (2004) utilisent une méthode de distance. La méthode de Tzeng (2005), quant à elle, ne nécessite pas la reconstruction d'un arbre. Seules les fréquences haplotypiques et les nombres de pas séparant les haplotypes sont importants pour établir les probabilités de regroupement. Les méthodes d'analyses sont aussi relativement différentes : Templeton *et al.* (1987, 2005) et Tzeng (2005) utilisent des ANOVA, des NANVOVA ou des tests de χ^2 selon les types de données, alors que Durrant *et al.* (2004) et Seltman *et al.* (2003) utilisent des modèles linéaires. Enfin, les méthodes de Templeton *et al.* (1987); Seltman *et al.* (2003) et Durrant *et al.* (2004) utilisent l'arbre pour réaliser une analyse emboîtée des données alors que Templeton *et al.* (2005) utilisent l'arbre pour définir toutes les comparaisons possibles entre deux groupes.

Certaines de ces méthodes possèdent des spécificités particulières :

- la méthode de Durrant *et al.* (2004) permet d'analyser de longues séquences et de localiser la région contenant le variant à risque sur le chromosome, même s'il n'est pas génotypé ;
- la méthode de Templeton *et al.* (2005) permet de prendre en compte l'information génotypique ;
- la méthode de Tzeng (2005) permet de s'affranchir complètement du problème des catégories rares car elles seront fusionnées avec des catégories plus fréquentes. Ce processus de regroupement peut cependant induire une perte d'information sur les allèles rares. Il ne permet pas non plus de faire des hypothèses sur la localisation du site de susceptibilité.

Enfin, hormis la méthode de Templeton *et al.* (1987), ces méthodes sont relativement récentes. Elles ont donc, pour l'instant, été assez peu utilisées et il y a très peu d'études évaluant leur puissance et la comparant à celles de méthodes plus classiques.

Chapitre 3

La méthode ALTree

Sommaire

3.1 Principe de la méthode ALTree . . . 62
3.1.1 Détection de l'association . . . 63
3.1.1.1 Principe du test . . . 63
3.1.1.2 Correction pour les tests multiples . . . 63
3.1.1.3 Gestion des arbres équiparcimonieux . . . 63
3.1.2 Localisation des sites de susceptibilité . . . 63
3.1.2.1 Principe de la méthode . . . 66
3.1.2.2 Gestion des arbres équiparcimonieux . . . 67
3.2 Implémentation de la méthode au sein du logiciel ALTree . . . 67
3.2.1 Résumé de l'article 1 : ALTree : Association detection and Localization of susceptibility sites using haplotype phylogenetic Trees . . . 67
3.2.2 Article 1 . . . 68
3.3 Étude de puissance . . . 71
3.3.1 Le processus de simulation . . . 71
3.3.1.1 Obtention d'un ensemble d'haplotypes . . . 71
3.3.1.2 Choix des sites de susceptibilité . . . 73
3.3.1.3 Tirage aléatoire des haplotypes et attribution d'un phénotype 73
3.3.1.4 Reconstruction phylogénétique . . . 73
3.3.1.5 Définition de la puissance de détection de l'association et de l'efficacité de localisation . . . 74
3.3.2 Résumé de l'article 2 : On the use of haplotype phylogeny to detect disease susceptibility loci . . . 75
3.3.3 Article 2 . . . 76
3.4 Les facteurs influençant la méthode . . . 91
3.4.1 La taille de l'échantillon . . . 91
3.4.2 Le modèle maladie . . . 92
3.4.2.1 La fréquence de l'allèle de susceptibilité . . . 92
3.4.2.2 La pénétrance de la maladie . . . 94
3.4.3 Les paramètres liés à la reconstruction phylogénétique . . . 96
3.4.3.1 La méthode de reconstruction de l'arbre . . . 96
3.4.3.2 Comparaison des méthodes de parcimonie et de vraisemblance pour l'optimisation des caractères . . . 98
3.4.3.3 L'enracinement et les méthodes d'optimisation des caractères 99

3.4.4 Variabilité liée aux jeux de données 102
3.5 Étude de puissance d'une méthode phylogénétique alternative 104
3.5.1 Nouvelle étude de puissance de la méthode de Durrant *et al.* 104
3.5.1.1 Résumé de l'article 3 : Clustering of haplotypes based on phylogeny : how good a strategy for association testing ? . 104
3.5.2 Article 3 106

Hormis les travaux de Templeton, les différentes méthodes phylogénétiques exposées dans le chapitre précédent ont été développées tout récemment, entre 2001 et 2005. Au moment où nous avons commencé notre étude, seuls les travaux de Templeton et de Seltman étaient publiés. Ces deux études ne nous satisfaisaient pas sur plusieurs points :

Test de la puissance La question de la puissance de détection de l'association n'est pas du tout traitée dans les articles de Templeton. Dans l'article de Seltman, quelques simulations sont effectuées, mais elles sont peu nombreuses (13 conditions) : peu de réplicats sont réalisés (200) et peu d'haplotypes sont considérés (11). De plus, le mode de simulation (choix de clades à risques, le statut maladie d'un individu étant choisi selon son appartenance à tel ou tel clade) semble favoriser les analyses phylogénétiques.

Pas de méthode de localisation des sites de susceptibilité Aucune des deux études ne propose de méthode systématique de localisation du site de susceptibilité. Seltman *et al.* (2001) proposent d'identifier les haplotypes à risque sans émettre plus d'hypothèses sur le ou les locus précis impliqués dans le déterminisme de la maladie. De leur côté, Templeton *et al.* (1987) proposent d'utiliser les haplotypes recombinants pour localiser la région contenant le ou les sites de susceptibilité. Or, sur des jeux de données plus conséquents que ceux qu'ils utilisent, il est souvent très difficile d'identifier les haplotypes recombinants.

Pas de logiciels disponibles pour réaliser ces analyses En 2002, aucun logiciel permettant de réaliser ce type d'analyse n'était disponible sur internet.

Nous avons donc développé une nouvelle méthode permettant de détecter une association entre un gène candidat et une maladie et d'identifier le locus de susceptibilité impliqué dans le déterminisme de cette maladie. Nous l'avons implémentée dans le logiciel `ALTree` .

3.1 Principe de la méthode `ALTree`

La méthode `ALTree` (**A**ssociation detection and **L**ocalization of susceptibility sites using haplotype phylogenetic **Tree**s) analyse des données haplotypiques formées de marqueurs biallélliques (SNPs, en particulier). Dans cette section, nous allons développer la méthode que nous avons mise au point pour détecter une association et pour localiser le ou les sites de susceptibilité à la maladie. Dans toute la section, nous allons exposer la méthode dans le cas où les phylogénies sont reconstruites par parcimonie ; mais des méthodes de maximum de vraisemblance peuvent parfaitement être utilisées comme nous le verrons au paragraphe 3.4.3.1.

3.1.1 Détection de l'association

3.1.1.1 Principe du test

L'association est détectée par des tests d'homogénéité emboîtés qui comparent les distributions des malades et des témoins dans les différents clades définis sur l'arbre phylogénétique. Au préalable, un arbre phylogénétique doit être reconstruit. L'analyse est réalisée sur un cladogramme, c'est-à-dire, un arbre dont seule la structure est pertinente et pas les longueurs des branches. Comme l'analyse emboîtée démarre au niveau de la racine, l'arbre doit être enraciné à partir d'une séquence ancestrale ou en utilisant un groupe externe.

Le principe des tests d'homogénéité emboîtés est expliqué sur la Figure 3.1. À chaque niveau, c'est-à-dire à chaque fois que de nouvelles bifurcations (ou multifurcations) sont définies sur l'arbre, on compare le nombre de malades et de témoins présents dans les différents clades définis à ce niveau. Les p-values des tests d'homogénéité sont calculées en utilisant les lois de distribution asymptotiques du χ^2 de Pearson sauf dans deux cas :

- lorsque seulement deux clades sont comparés, les p-values sont évaluées en utilisant les lois de distribution asymptotiques du χ^2 exact de Fisher ;
- lorsque plus de deux clades sont comparés et que les effectifs dans au moins un des clades sont faibles (moins de 4 haplotypes), les p-values sont évaluées par permutations.

3.1.1.2 Correction pour les tests multiples

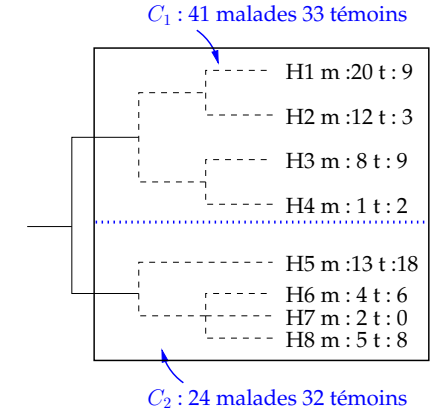
Pour mettre en évidence une association, on réalise un test par niveau de l'arbre. Lorsque l'arbre est complexe, le nombre de tests peut être relativement grand. Il est alors nécessaire de réaliser une correction pour les tests multiples afin d'évaluer le degré réel de signification des différents tests d'homogénéité. Comme ces tests sont emboîtés, ils sont extrêmement dépendants entre eux. Une correction de Bonferroni n'est donc pas du tout appropriée car elle va être trop conservatrice et va diminuer énormément la puissance du test. Dans `ALTree` , nous avons implémenté une procédure de permutation identique à celle décrite par Ge *et al.* (2003) (voir page 20), bien adaptée au traitement des tests emboîtés, ce qui est le cas sur l'arbre phylogénétique.

3.1.1.3 Gestion des arbres équiparcimonieux

Pour un ensemble d'haplotypes donné, on obtient généralement plusieurs arbres équiparcimonieux. L'idéal serait de pouvoir tous les analyser. Cependant, la procédure de permutation de Ge *et al.* (2003) est très longue (2 jours sur un Pentium III, 930 MHz, 512 Mo de RAM). On se contentera alors généralement de réaliser le test d'association sur l'un des arbres équiparcimonieux tiré aléatoirement.

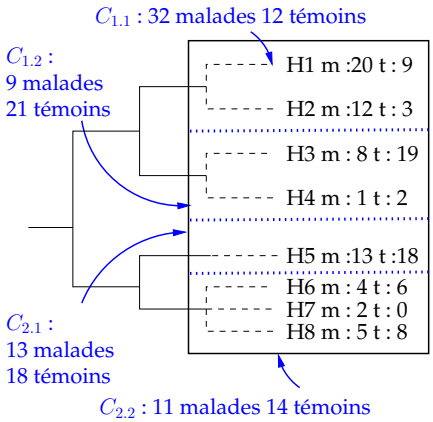
3.1.2 Localisation des sites de susceptibilité

Au cours de mon DEA et au début de ma thèse, nous avons travaillé sur une première méthode de localisation des sites de susceptibilité (Bardel, 2002). Cette méthode est basée sur la réalisation de tests d'homogénéité de distribution des malades et des témoins, et de tests de proportion permettant d'identifier les clades comportant un excès significatif de malades. Les synapomorphies définissant ces clades sont les sites putatifs de susceptibilité



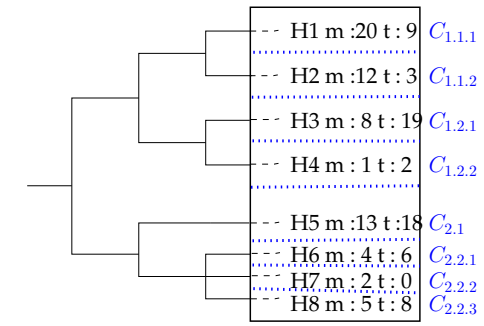
(a) Niveau 1

Au niveau 1 (première division dans l'arbre), un test d'homogénéité de la distribution des malades et des témoins entre les clades C_1 et C_2 est réalisé (1 degré de liberté). La p-value de ce test est de 0,16.



(b) Niveau 2

Au niveau 2, un test d'homogénéité de la distribution des malades et des témoins entre les clades $C_{1.1}$, $C_{1.2}$, $C_{2.1}$ et $C_{2.2}$ — qui sont les descendants des clades C_1 and C_2 — est réalisé (3 degrés de liberté). La p-value de ce test est de 0,0019.



(c) Niveau 3

Au niveau 3, un test d'homogénéité de la distribution des malades et des témoins entre les clades $C_{1.1.1}$, $C_{1.1.2}$, $C_{1.2.1}$, $C_{1.2.2}$ $C_{2.1}$, $C_{2.2.1}$, $C_{2.2.2}$ et $C_{2.2.3}$ est réalisé (7 degrés de liberté). La p-value de ce test est de 0,057.

FIG. 3.1: Description de l'analyse emboîtée lors du test d'association

Les figures 3.1a, 3.1b et 3.1c représentent les 3 niveaux de l'analyse emboîtée sur cet arbre. Pour chaque catégorie haplotypique (H1 à H8), le nombre d'haplotypes porté par des individus malades est indiqué après la lettre m et le nombre d'haplotypes portés par des individus témoins est indiqué après la lettre t. Sur chaque figure, les différents clades sont séparés par des pointillés.

On peut remarquer qu'au niveau 2, le clade $C_{2.1}$ correspond en fait à la catégorie haplotypique H5. Cette branche pourrait s'arrêter là et ne plus être prise en compte par la suite. Nous avons cependant choisi de "prolonger" cette branche et de tenir compte de ce clade au niveau 3, afin de conserver le même échantillon d'individus à tous les niveaux.

à la maladie. Le principe de cette méthode est détaillé en annexe A. Cette méthode bien que relativement puissante (Bardel, 2002) présente trois principales limites :

1. Les tests — significatifs ou non — réalisés au niveau k dépendent des résultats de ceux réalisés au niveau $k-1$. Il faut donc absolument fixer a priori des seuils de significativité pour les tests d'homogénéité qui tiennent compte des tests multiples réalisés. Or, il est très difficile de fixer ce seuil car on ne peut utiliser ni permutation ni correction de Bonferroni. En effet, pour cette dernière, il nous faudrait connaître le nombre de tests a priori, ce qui n'est pas le cas. De plus, les tests n'étant pas indépendants, elle serait de toute façon trop conservatrice ;
2. Cette méthode ne permet de traiter qu'un seul arbre à la fois. Or, les arbres équiparcimonieux peuvent être nombreux et raconter des histoires plus ou moins différentes. Afin d'obtenir une meilleure efficacité, il apparaît important de tenir compte de l'ensemble de ces arbres. Ce problème aurait pu cependant être relativement facilement résolu en intégrant les résultats obtenus sur tous les arbres équiparcimonieux ;
3. Le dernier problème concerne les petits échantillons. Lorsque des haplotypes sont portés par un très faible nombre d'individus, les tests statistiques ne sont pas assez puissants pour localiser correctement le site de susceptibilité. La Figure 3.2 présente un exemple de ce type de problème obtenu au cours de nos simulations.

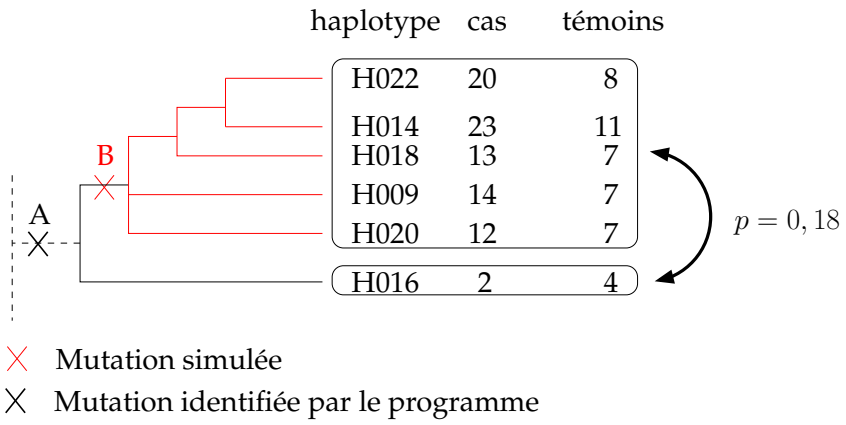


FIG. 3.2: Erreur de localisation du fait des petits échantillons

Le test d'homogénéité de distribution des malades et des témoins entre le clade B et l'haplotype H016 n'est pas significatif ($p = 0.18$). Le programme va donc considérer à tort que l'ensemble du clade A contient un excès de malades. Le site de susceptibilité ne pourra donc pas être détecté.

En raison de toutes ces limites, nous avons abandonné ce type de méthode de localisation et nous en avons élaboré une autre, davantage basée sur des concepts évolutifs et sur les propriétés de l'arbre phylogénétique. De plus, cette nouvelle méthode permet de s'affranchir de l'enracinement, ce qui est particulièrement intéressant pour les analyses de données où la séquence ancestrale n'est pas connue. Ainsi, nous n'avons pas besoin de faire d'hypothèses concernant l'ancêtre.

3.1.2.1 Principe de la méthode

Elle consiste à définir pour chaque haplotype un nouveau caractère dont l'état dépend du statut — malade ou témoin — des individus portant cet haplotype. Les états des caractères aux différents nœuds de l'arbre sont ensuite inférés, ce qui permet d'identifier les branches sur lesquelles les sites changent d'état. Pour chaque site ainsi que pour chaque sens de changement d'état de ce site, on définit un indice décrivant l'évolution corrélée entre le site et le nouveau caractère. Le/les sites ayant l'indice le plus élevé sont les sites putatifs de susceptibilité à la maladie.

Définition du nouveau caractère Pour chaque haplotype, ce nouveau caractère, nommé S, est défini. Les états de S sont "T" (témoin), "M" (malade) ou " ?" (statut indéterminé) selon la proportion p_h de malades portant l'haplotype h par rapport à la proportion p_0 de malades dans l'ensemble de l'échantillon. Le caractère S pourrait être défini de différentes manières. Nous proposons de le faire de la façon suivante :

- si $p_h < p_0 - \epsilon\sqrt{\frac{p_h \times (1-p_h)}{n_h}}$, S est codé "T" (haplotype majoritairement porté par des témoins) ;
- si $p_h > p_0 + \epsilon\sqrt{\frac{p_h \times (1-p_h)}{n_h}}$, S est codé "M" (haplotype majoritairement porté par des malades) ;
- sinon, S est codé " ?" (donnée indéterminée).

n_h étant le nombre d'individus portant l'haplotype h et ϵ une valeur seuil choisie par l'utilisateur de la méthode (généralement, nous prenons $\epsilon = 1$).

Optimisation de S sur l'arbre phylogénétique Ce nouveau caractère S n'est pas utilisé pour reconstruire la phylogénie des haplotypes. En revanche, ses états aux différents nœuds de l'arbre sont reconstruits en utilisant les même options que pour les autres sites. Cela permet d'identifier les branches sur lesquelles ce caractère change d'état.

Définition de l'indice de co-évolution V_i Lors de la définition de l'indice de co-évolution V_i, nous supposons que les mutations sont distribuées de façon aléatoire sur l'arbre et que chaque branche a la même probabilité de porter une mutation.

Cet indice compare le nombre observé de co-mutations entre le caractère S et le site i avec le nombre attendu de co-mutations si les mutations sont distribuées aléatoirement. Il est défini pour un arbre t, pour chaque site i pouvant présenter les états 0 et 1, et pour chaque sens de changement d'état : $0 \to 1$ et $1 \to 0$. Par conséquent, nous noterons $V_{i,t}^{0\to1}$ l'indice de co-évolution entre la transition $0 \to 1$ du site i et le caractère S sur l'arbre t.

L'indice de co-évolution $V_{i,t}^{0\to1}$ (resp. $V_{i,t}^{1\to0}$) est défini de la façon suivante :

$$\begin{cases} V_{i,t}^{0\to1} = 0 & si\ E_{i,t}^{0\to1} = 0 \\ V_{i,t}^{0\to1} = \dfrac{R_{i,t}^{0\to1} - E_{i,t}^{0\to1}}{\sqrt{E_{i,t}^{0\to1}}} & si\ E_{i,t}^{0\to1} \neq 0 \end{cases}$$

où,

- $R_{i,t}^{0\to1}$ (resp. $R_{i,t}^{1\to0}$) est le nombre de co-mutations observées sur l'arbre entre le site i et le caractère S

– $E_{i,t}^{0\to1}$ (resp. $E_{i,t}^{1\to0}$) est le nombre attendu de co-mutations entre le caractère S et le site i. $E_{i,t}^{0\to1}$ est calculé de la façon suivante[1] :

$$E_{i,t}^{0\to1} = \frac{(m_{i,t}^{0\to1} \times s_t^{T\to M}) + (m_{i,t}^{1\to0} \times s, t^{M\to T})}{b_t}$$

où :

$m_{i,t}^{0\to1}$ (resp. $m_{i,t}^{1\to0}$) est le nombre de transitions $0 \to 1$ (resp. $1 \to 0$) du caractère i ;

$s_t^{T\to M}$ (resp. $s_t^{M\to T}$) est le nombre de transitions $T \to M$ (resp. $M \to T$) du caractère S ;

b_t est le nombre de branches de l'arbre t.

3.1.2.2 Gestion des arbres équiparcimonieux

Afin de considérer l'ensemble des arbres équiparcimonieux, on définit ensuite un indice de co-évolution pour l'ensemble des T arbres équiparcimonieux en sommant les $V_{i,t}^{0\to1}$ (resp. $V_{i,t}^{1\to0}$) :

$$V_i^{0\to1} = \sum_{t=1}^{T} V_{i,t}^{0\to1}$$

3.2 Implémentation de la méthode au sein du logiciel ALTree

Nous avons implémenté notre méthode phylogénétique dans le logiciel ALTree (environ 5000 lignes de code en perl). Ce logiciel, ainsi que sa documentation, est disponible sur le web à l'adresse suivante : http://claire.bardel.free.fr.

3.2.1 Résumé de l'article 1 : ALTree : Association detection and Localization of susceptibility sites using haplotype phylogenetic Trees

Le logiciel ALTree est composé de trois programmes : ALTree , le programme principal correspondant à l'implémentation de la méthode décrite dans les paragraphes précédents, et deux utilitaires, ALTree-convert et ALTree-add-S qui aident l'utilisateur à mettre en forme ses données avant d'utiliser ALTree .

ALTree : Ce programme permet soit de détecter une association, soit de localiser des sites de susceptibilité suivant l'option choisie par l'utilisateur. Il nécessite deux fichiers d'entrée, l'un contenant la liste des haplotypes et le nombre de malades et de témoins correspondant, et le second contenant l'arbre phylogénétique et la liste des changements d'état des caractères. Ce dernier fichier est généré par un logiciel de reconstruction phylogénétique, soit PAUP (Swofford, 2002) ou PHYLIP (Felsenstein, 2004c), pour une analyse par parcimonie, soit PAML (Yang, 1997) pour une analyse par maximum de vraisemblance.

[1] Dans l'article 2, nous n'avons pas exactement utilisé cette manière de calculer $E_{i,t}^{0\to1}$. À ce moment là, nous n'avions pas encore choisi de prendre en compte les changements $M \to T$ du caractère S. C'est pourquoi nous avions :

$$E_{i,t}^{0\to1} = \frac{(m_{i,t}^{0\to1} \times s_t^{T\to M})}{b_t}$$

ALTree -convert : Pour des individus diploïdes, les haplotypes ne sont pas directement observables. Le séquençage permet seulement d'avoir accès aux génotypes non phasés. Lors d'une analyse de données, les haplotypes sont alors reconstruits en utilisant des programmes spécifiques. `ALTree-convert` permet de convertir les fichiers de sortie de deux programmes de reconstruction haplotypique, `FAMHAP` (Becker et Knapp, 2004b) et `PHASE` (Stephens *et al.*, 2001a; Stephens et Donnelly, 2003) en fichiers d'entrée utilisables directement par les logiciels de reconstruction phylogénétique (format `PAUP` ou `PHYLIP`). Dans le cas du format `PAUP` , le fichier obtenu par `ALTree-convert` contient une liste de commandes qui permettent de faire fonctionner le logiciel `PAUP` . Seules quelques modifications mineures du fichier (ajout de la séquence ancêtre, choix du nombre maximum d'arbres recherchés, noms des fichiers de sortie, etc.) doivent être faites au préalable. Cela permet à des utilisateurs peu familiers de ce logiciel de l'utiliser facilement.

ALTree-add-S : La localisation du site de susceptibilité nécessite de définir les états du caractère S pour chaque haplotype. Pour cela, l'utilisateur peut modifier manuellement le fichier d'entrée de `PAUP` pour ajouter ce caractère selon ses propres critères. S'il souhaite le faire de façon automatique, il peut utiliser le programme `ALTree-add-S` . Ce programme prend en entrée un fichier format `PAUP` ou `PHYLIP` ainsi qu'une liste d'haplotypes et le nombre de malades et de témoins correspondant. Il donne en sortie un nouveau fichier `PAUP` ou `PHYLIP` dans lequel le caractère S a été ajouté à tous les haplotypes selon le principe décrit au page 66 .

Le logiciel `ALTree` peut tourner sous différents systèmes d'exploitation (linux, Mac OSX). Le logiciel étant écrit en perl, avec un module en C (évaluation des p-values pour le test d'association), il est nécessaire que la version 5.8 de perl ainsi qu'un compilateur C soient installés sur le système.

3.2.2 Article 1

ALTree : Association detection and Localization of susceptibility sites using haplotype phylogenetic Trees (Bardel *et al.*, 2006)
Claire Bardel, Vincent Danjean et Emmanuelle Génin. Soumis à **Bioinformatics**.

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

ALTree: Association detection and Localization of susceptibility sites using haplotype Phylogenetic Trees

Claire Bardel [a]*, Vincent Danjean [b] and Emmanuelle Génin[a]

[a]Unité de recherche en Génétique Épidémiologique et Structure des Populations Humaines, INSERM U535, Villejuif, France, [b] Laboratoire Bordelais de Recherche en Informatique, UMR 5800, Bordeaux, France

ABSTRACT

Summary:

Finding the genes involved in complex disease susceptibility and among those genes, localizing the variant sites explaining this susceptibility is a major goal of genetic epidemiology. In this context, haplotypic methods that use the joint information on several markers may be of particular interest. When the number of haplotypes is large, a grouping may be required. Phylogenetic trees allow such groupings of haplotypes based on their evolutionary history and may help in the detection and localization of disease susceptibility sites. In this paper, we present a new software to perform phylogeny-based association and localization analyzes.

Availability: The software package, including all documentation and example is freely available at `http://claire.bardel.free.fr`. It is distributed under the GPL license.

Contact: bardel@vjf.inserm.fr

1 INTRODUCTION

When looking for an association between a candidate gene and a disease, genetic markers can be either studied one at a time or combined with the other markers located on the same chromosome to form haplotypes. In the last few years, haplotypic methods have been shown to be powerful to look for association (Akey *et al.*, 2001; Zaykin *et al.*, 2002). However, when the number of haplotypes increases, the power of the association test will decrease, due to the increase in the degree of freedom and to the decrease in the sample size per haplotype. To reduce these problems, it has been proposed to group haplotypes according to their evolutionary history (Templeton *et al.*, 1987; Seltman *et al.*, 2003; Durrant *et al.*, 2004; Bardel *et al.*, 2005). Moreover, such groupings allows to make hypotheses about the markers responsible for the susceptibility to the disease: in the evolutionary tree, mutations defining a group containing more haplotypes carried by cases than controls are putative susceptibility sites for the disease.

In this paper, we present a new software to perform a phylogeny-based association and localization analysis based on the method described in Bardel *et al.* (2005). The software deals with SNP haplotype data. It is written in perl, with one package in C. It is composed of three programs: the main analysis program, `altree`, which performs either the association or the localization analysis depending on the option selected by the user and two utilities, `altree-add-S` and `altree-convert`, which help the user managing its data file before running `altree`.

*to whom correspondence should be addressed

2 THE MAIN PROGRAM: `ALTREE`

To run `altree`, two input files are required: first, a file containing for each haplotype, the number of case and control individuals carrying it, and second, a file containing the phylogenetic tree and the list of character state changes on the tree. This last file is obtained by running a phylogeny reconstruction program. Currently, `altree` can deal with three phylogeny softwares: `paup` (Swofford, 2002) and `phylip` (Felsenstein, 2004) for a cladistic tree reconstruction, and `paml` (Yang, 1997) for a maximum likelihood tree reconstruction.

2.1 Association detection

The principle of the method is to perform series of nested case/control homogeneity tests in the different clades defined on the phylogenetic tree. P-values are computed at each level of the tree and a global p-value corrected for multiple testing is computed for the test, using a double permutation procedure (Becker and Knapp, 2004). The association test required the tree to be rooted because the nested analysis began at the root of the tree. Consequently, in the data set, an outgroup or an ancestral haplotype must be provided by the user so that `altree` can root the tree.

The output file contains the phylogenetic tree, with the number of case and control haplotypes in each branch, and the p-value of the association test.

2.2 Localization of the susceptibility site(s)

The principle of the localization analysis is to use a new character called "S" that represents the disease status. For each haplotype the state of this character depends on the proportion of cases carrying this haplotype (state 1 for a large proportion of cases and 0 else). The character state changes are optimized on the tree for each character (including S), and a correlated evolution index is calculated between each changes of each site and the changes of the character S. The site(s) whose evolution is the most correlated to the S character, is the most probable susceptibility site.

For the localization test, the character states can be reconstructed on the tree either by `paup` (parsimony method) or `paml` (Maximum likelihood method), but not by `phylip` because it leaves too many ambiguities in the ancestral character reconstructions. Contrary to the association analysis, the tree does not need to be rooted. When several trees are available (for example, several equiparsimonious trees), the user can specify how many trees should be included in the analysis. They will then be picked up at random from the total sample of parsimonious trees. Two kinds of correlated evolution index can be calculated. The first is the one described in Bardel

© Oxford University Press 2005.

et al. (2005): only transitions from 0 to 1 for the character S are taken into account. A second index which seems more appropriate is now available: it takes into account both transitions of S (from 0 to 1 and from 1 to 0). The correlated evolution index can be chosen using the option `--coevo`

The output file of the program consists in a list of all the correlated evolution index for all transitions of all sites, ranked in decreasing order. The phylogenetic trees can be included in the output file using the option `--print-tree`.

3 TWO UTILITIES

3.1 A file converter: `altree-convert`

Our software analyzes haplotypic data. Such data are not generally directly available and they must be obtained by using haplotype reconstruction programs. `altree-convert` allows to convert output files from two haplotype reconstruction programs, `phase` (Stephens *et al.*, 2001; Stephens and Donnelly, 2003) and `FamHap` (Becker and Knapp (2004), only for haplotype reconstructed on family data) into input files for the phylogeny reconstruction programs (`paup` or `phylip` file format). When the `paup` file format is chosen, a list of commands is also written in the file so that users who are not familiar with `paup` can run it with only a few modifications to this file.

3.2 Definition of the S character: `altree-add-S`

To define the state of the character S, the user can choose its own criterion and add the character S manually in the `paup` input file. Otherwise, the user can use `altree-add-S`. This program takes a `paup` or a `phylip` input file and a list of the number of cases and controls carrying a given haplotype in input and returns a new `paup` or `phylip` input file in which the S character has been added according to the following criterion: the state of S is "0", "1" or "?" depending on the proportion (p_h) of cases carrying the haplotype h compared to the proportion p_0 of cases in the whole sample.

- if $p_h < p_0 - \epsilon\sqrt{\frac{p_h \times (1-p_h)}{n_h}}$, S is coded "0" (high number of controls);
- if $p_h > p_0 + \epsilon\sqrt{\frac{p_h \times (1-p_h)}{n_h}}$, S is coded "1" (high number of cases);
- else, S is coded "?" (missing data).

with n_h being the number of individuals carrying the haplotype h and ϵ, a number chosen by the user.

4 REQUIREMENTS

The software requires perl 5.8, a C compiler and a phylogeny reconstruction software: `paup`, `phylip` or `paml`.

5 CONCLUSION

Our software groups haplotypes based on their phylogenetic relationships to perform both association and localization analyzes. With the two utilities, it is easily usable, even for users who are not accustomed to phylogeny reconstruction programs. The efficiency of the method implemented in this software has been evaluated by simulations and on a data set concerning Crohn disease. We have shown that it is especially interesting when more than one susceptibility site is involved in the disease (Bardel *et al.*, 2005).

REFERENCES

Akey, J., Jin, L. and Xiong, M. (2001) Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? *European J of Hum Genet*, **9**, 291–300.

Bardel, C., Danjean, V., Hugot, J. P., Darlu, P. and Génin, E. (2005) On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genetics*, **6**.

Becker, T. and Knapp, M. (2004) A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet*, **75**, 561–570.

Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P. and Morris, A. P. (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotique polymorphism haplotypes. *Am J Hum Genet*, **75**, 35–43.

Felsenstein, J. (2004) Phylip (phylogeny inference package) version 3.6. `http://evolution.genetics.washington.edu/phylip.html`. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Seltman, H., Roeder, K. and Devlin, B. (2003) Evolutionary-based association using haplotype data. *Genet Epidemiol*, **25**, 48–58.

Stephens, M. and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, **73**, 1162–1169.

Stephens, M., Smith, N. J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, **68**, 978–989.

Swofford, D. L. (2002) `paup` phylogenetic analysis using parcimony. version 4.0b10. Sunderland, Massachusetts: Sinauer Associates.

Templeton, A. R., Boerwinkle, E. and Sing, C. F. (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics*, **117**, 343–351.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer application in bioSciences*, **13**, 555–556. Http://abacus.gene.ucl.ac.uk/software/paml.html.

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

3.3 Étude de puissance

Nous avons ensuite étudié la puissance de la méthode implémentée dans `ALTree` pour détecter l'association, ainsi que son efficacité pour localiser les sites de susceptibilité en réalisant des simulations.

3.3.1 Le processus de simulation

Le processus général de simulation est représenté sur la Figure 3.3.

3.3.1.1 Obtention d'un ensemble d'haplotypes

Pour obtenir un ensemble d'haplotypes qui vont servir à constituer des génotypes d'individus malades ou témoins, nous avons soit réalisé des simulations par une méthode basée sur la coalescence, soit utilisé des données disponibles sur internet.

Simulation d'haplotypes sous un modèle de coalescence Nous avons simulé nos haplotypes en utilisant le logiciel `TREEVOLVE` . Il génère tout d'abord aléatoirement une séquence ancestrale et un arbre-guide sous un modèle de coalescence sans recombinaison. Ensuite, l'évolution de cette séquence ancestrale le long des branches de l'arbre est simulée en tenant compte des paramètres du modèle évolutif spécifié par l'utilisateur.

`TREEVOLVE` a été conçu pour simuler des séquences d'ADN (4 nucléotides). Or, nous souhaitons simuler des marqueurs biallélicques. Nous avons donc choisi des valeurs très particulières pour les différents paramètres :

- très faibles fréquences initiales de A et T ($A = T = 0.001$) et fréquences plus fortes de G et C ($G = C = 0.498$) de manière à ce que la séquence ancestrale soit formée uniquement de G et de C ;
- taux de substitution fixé de telle sorte que seuls les changements $C \longleftrightarrow G$ sont autorisés : $A - C = 0.001$, $A - G = 0.001$, $A - T = 0.001$, $C - T = 0.001$, $G - T = 0.001$ et $G - C = 0.995$;
- taux de mutation fixé à 2×10^{-6} ;
- pas d'hétérogénéité dans les taux de substitution entre les sites ;
- la taille de la population est fixée à 100 000 et sa croissance est stationnaire sans subdivision.

Utilisation d'haplotypes décrits dans une base de données Nous avons utilisé deux bases de données :

SeattlesSNPs Cette base de données (SeattleSNPs) regroupe des informations concernant de nombreux gènes séquencés chez 47 individus (23 européens et 24 africains). Sur ces gènes, les SNPs sont identifiés et ceux dont la fréquence de l'allèle mineur est supérieure à 5% dans au moins une des deux populations sont utilisés pour reconstruire les haplotypes des 47 individus en utilisant le logiciel `PHASE` (Stephens *et al.*, 2001a; Stephens et Donnelly, 2003). Dans nos simulations, nous n'avons utilisé que les haplotypes reconstruits des individus de la population européenne.

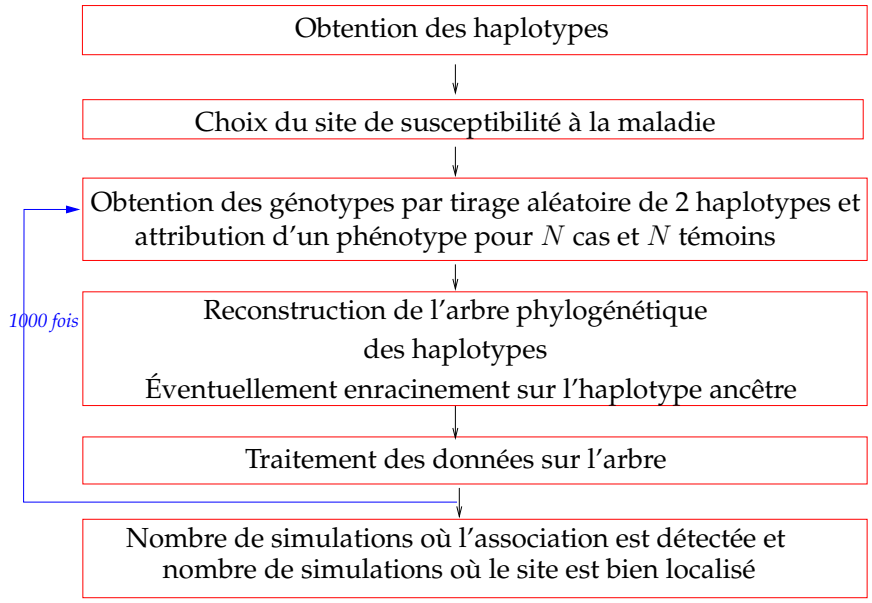


FIG. 3.3: Les différentes étapes d'une simulation

Gène	Taille (bp)[e]	Chromosome	Nb SNPs	Nb haplotypes[a]
PLAU[b]	9274	10q24	16	10
IL13[c]	6919	5q31	12	14
TNF[d]	4830	6p21.3	10	6

[a] Il s'agit du nombre d'haplotypes différents reconstruits par PHASE dans la population européenne uniquement ;
[b] *Plasminogen Activator Urokinase*
[c] *Interleukine 13*
[d] *Tumor Necrosis Factor*
[e] La taille des séquences est exprimée en paires de bases

TAB. 3.1: Caractéristiques des 3 gènes de la base Seattles SNPs

Dans cette base de données, nous avons utilisé les données concernant 3 gènes dont les caractéristiques sont récapitulées dans le Tableau 3.1.

Il s'agit donc de régions courtes avec un marqueur toutes les 500 à 600 paires de bases.

Données du projet HapMap Le but du projet HapMap est de fournir un catalogue des polymorphismes courants chez l'homme avec leurs descriptions, leurs localisations sur les chromosomes et leurs fréquences dans 4 populations (africaine, japonaise, chinoise et américaine originaire d'Europe de l'ouest).

Dans cette base de données, nous avons sélectionné une région de 500 kb sur le chromosome 20 dans laquelle nous avons choisi aléatoirement 13 et 12 polymorphismes pour former les données nommées CHR20_1 et CHR20_2 respectivement. Nous avons utilisé les données génotypiques fournies pour 30 trios (père, mère, enfant) de la population américaine. Les

haplotypes des individus ont été reconstruits par le logiciel PHASE [2]

Nous avons obtenu les deux jeux de données décrits dans le Tableau 3.2 Il s'agit donc

Gène	Taille (kb)[b]	Nb SNPs	Nb haplotypes[a]
CHR20_1	461	13	37
CHR20_2	442	12	38

[a] Il s'agit du nombre d'haplotypes différents reconstruits par PHASE ;
[b] La taille des séquences est exprimée en kilo paires de bases

TAB. 3.2: Caractéristiques des deux régions du chromosome 20

de séquences plus longues que celles des données de la base (SeattleSNPs), avec environ un marqueur tous les 35 kb.

Les fréquences alléliques des marqueurs et leur localisation dans les différentes régions du génome sont présentées dans l'annexe B.

3.3.1.2 Choix des sites de susceptibilité

Les sites de susceptibilité sont choisis en fonction de leurs fréquences alléliques. Pour un modèle de maladie où un seul allèle de fréquence f est à risque, on choisit comme site de susceptibilité le premier site sur la séquence étudiée, dont la fréquence d'un des allèles vaut f. Lorsque plusieurs sites en interaction sont responsables de la susceptibilité à la maladie, le principe est le même, mais ce sont les fréquences des haplotypes formés par les allèles des différents sites qui vont être prises en compte.

3.3.1.3 Tirage aléatoire des haplotypes et attribution d'un phénotype

Les génotypes des individus sont ensuite formés par tirage aléatoire avec remise de deux haplotypes. Les statuts maladie des individus sont attribués en fonction de leurs génotypes au locus de susceptibilité et des pénétrances.

Ce processus est répété jusqu'à former un échantillon de N malades et un échantillon de N témoins, ce qui constitue alors un *réplicat*.

3.3.1.4 Reconstruction phylogénétique

La phylogénie des haplotypes est reconstruite soit par une méthode de parcimonie, soit par une méthode de maximum de vraisemblance.

Reconstruction en utilisant une méthode de parcimonie Nous avons choisi d'utiliser la méthode de parcimonie implémentée dans le logiciel PAUP (version 4.0b10, Swofford

[2]Le logiciel PHASE n'est pas très adapté sur des données trios car il ne prend pas en compte les relations de parenté : les haplotypes des enfants et ceux des parents sont reconstruits indépendamment. Comme nous souhaitions juste obtenir un ensemble d'haplotypes pour réaliser nos simulations, nous nous sommes contentés de cette approximation et nous n'avons pas recherché les relations de parenté entre les individus (elles ne sont pas spécifiées dans les fichiers HapMap)

(2002)). Nous avons réalisé une recherche heuristique des arbres parcimonieux en utilisant un algorithme d'addition pas à pas des haplotypes et le réarrangement de branches TBR (*tree bisection and reconnection*). Nous avons donné le même poids[3] à tous les sites et tous les caractères sont non ordonnés, c'est-à-dire que tous les sens de changement d'état sont autorisés pour tous les sites (site de susceptibilité inclus).

Pour les analyses d'association et pour certaines analyses de localisation, nous avons enraciné l'arbre en spécifiant une séquence ancestrale, ce qui permet de définir pour chaque caractère un état ancestral et un état dérivé de façon *a priori*[4]. Les états de caractère aux nœuds sont inférés en utilisant les options `acctran` ou `deltran`.

Reconstruction en utilisant une méthode de maximum de vraisemblance Nous avons choisi d'utiliser la méthode de maximum de vraisemblance (ML) implémentée dans le logiciel `PHYML` (Guindon et Gascuel, 2003) car il est le seul à pouvoir traiter des jeux de données contenant plus d'une dizaine de taxons dans un temps très court. Il peut donc être utilisé lors de simulations, ce qui est totalement impossible avec les autres logiciels actuellement disponibles. L'arbre initial est obtenu par *neighbor-joining* (NJ). Nous avons choisi d'utiliser le modèle de substitution REV[5] (aussi nommé GTR) car il nous permet de modéliser l'évolution de SNPs bialléliques. Nous avons aussi fixé le nombre de sites invariants à 0 et nous avons supposé que le taux de substitution est le même pour tous les sites.

Le logiciel `PHYML` ne donne que l'arbre le plus vraisemblable, c'est pourquoi nous n'avons analysé qu'un seul arbre par ML alors que nous avons conservé tous les arbres équiparcimonieux. `PHYML` ne permet pas non plus de calculer les probabilités des états des différents caractères aux nœuds. Aussi avons-nous utilisé un second logiciel, `PAML` (Yang, 1997) pour réaliser cette étape. Les probabilités que nous avons utilisées sont les probabilités *a posteriori* des états des caractères estimées séparément pour chaque nœud (voir page 33).

Les méthodes de ML ne permettent pas d'enraciner l'arbre en polarisant les caractères *a priori*. L'enracinement étant nécessaire pour les études d'association, nous avons alors utilisé la séquence ancestrale comme groupe externe. Pour les analyses de localisation, nous avons choisi de ne pas enraciner l'arbre.

3.3.1.5 Définition de la puissance de détection de l'association et de l'efficacité de localisation

Les tests d'association et de localisation du site de susceptibilité ont été réalisés sur 1 000 réplicats simulés de la manière décrite dans les paragraphes précédents. Cela nous permet d'évaluer la puissance du test d'association et l'efficacité de la méthode de localisation du site de susceptibilité.

Puissance du test d'association La puissance du test d'association correspond à la probabilité de rejeter l'hypothèse nulle H_0 d'absence d'association sachant que l'hypothèse alternative (présence d'une association) est vraie. En pratique, nous la calculons ainsi :

[3] La pondération permet de donner une plus ou moins grande importance à un caractère lors de la reconstruction phylogénétique. Ici, tous les caractères auront la même importance.

[4] Il faut noter que cette hypothèse n'est pas contradictoire avec celle d'absence d'ordre : on observe aussi bien des changements d'état dans le sens ancestral → dérivé que dans le sens dérivé → ancestral

[5] Le modèle REV comporte 6 paramètres correspondant aux probabilités des changements d'état des 4 bases et il n'y a aucune contrainte sur les fréquences initiales des quatre bases

$$\text{Puissance} = \frac{\text{nombre de réplicats où le test d'association est significatif (rejet de } H_0)}{\text{nombre total de réplicats}}$$

Efficacité de la méthode de localisation Nous avons défini l'efficacité de la méthode de localisation comme étant la proportion de réplicats dans lesquels le site de susceptibilité simulé est identifié, soit de façon unique (dans `ALTree` , cela signifie que le site simulé possède le plus grand V_i), soit parmi un ensemble d'autres sites (dans `ALTree` , cela signifie que la valeur maximale du V_i est obtenue pour plusieurs sites, dont le site de susceptibilité). En pratique, nous la calculons ainsi :

$$\text{Efficacité} = \frac{\text{nombre de réplicats dans lesquels le site de susceptibilité est identifié}}{\text{nombre total de réplicats}}$$

Lorsque 2 sites de susceptibilité sont simulés, on définira :
– l'efficacité de détection des deux sites ;
– l'efficacité de détection d'au moins un des deux sites.

La puissance de détection de l'association et l'efficacité de la localisation du site de susceptibilité d'`ALTree` ont été étudiées et comparées à celles d'autres méthodes dans l'article 2.

3.3.2 Résumé de l'article 2 : On the use of haplotype phylogeny to detect disease susceptibility loci

L'article 2 présente la méthode de détection de l'association et de localisation des sites de susceptibilité décrite aux sections 3.1.1 et 3.1.2. La puissance et l'efficacité de cette méthode sont étudiées par simulations pour trois modèles génétiques différents et sont comparées à celles obtenues par d'autres méthodes. La puissance de détection de l'association est comparée à un test allélique locus par locus, à un test haplotypique sans groupement des haplotypes et au test phylogénétique implémenté dans `CLADHC` . Pour l'efficacité de localisation, seul le test locus par locus est comparé au test implémenté dans `ALTree` . Enfin, la méthode est appliquée à des données concernant la maladie de Crohn (ce dernier point sera développé dans le chapitre 4).

Des données sont simulées avec le logiciel `TREEVOLVE` (Grassly et Rambaut, 2000) pour 10 topologies d'arbres différentes et sont analysés selon la méthode décrite page 71. Chaque jeu de données est analysé sous trois modèles de maladie :

modèle 1 : Un seul site de susceptibilité, fréquence de l'allèle à risque (d) égale à $0,2$, pénétrances égales à $P_{DD} = 0,03$, $P_{Dd} = 0,06$, $P_{dd} = 0,30$;

modèle 2 : Deux sites de susceptibilité A_2 et B_2 en interaction ; fréquence de l'haplotype à risque A_2B_2 égal à $0,25$, pénétrance associée au génotype le plus à risque égale à $0,9$, toutes les autres pénétrances sont fixées à $0,3$;

modèle 3 : Mêmes valeurs que le modèle 2, mais avec la pénétrance associée au génotype le plus à risque égale à $0,6$.

Le calcul des p-values associées au test d'association est beaucoup trop long[6] pour être réalisé en simulations. Afin de faciliter les comparaisons des puissances des différents tests, nous avons donc contrôlé *a posteriori* que l'erreur de type I est bien autour de 5% pour toutes les méthodes en simulant 1000 réplicats sous l'hypothèse nulle d'absence d'association et en les analysant.

Nos simulations nous ont permis de montrer que lorsque le ou les sites de susceptibilité sont des marqueurs inclus dans l'analyse, les tests d'association phylogénétiques (`CLADHC` et `ALTree`) sont moins puissants que le test allélique locus par locus. En revanche, lorsque le site de susceptibilité ne fait pas partie des marqueurs génotypés, les tests phylogénétiques et en particulier `ALTree` permettent de gagner un peu de puissance. Cependant, il faut bien noter que les puissances des méthodes sont très proches, et comme les erreurs de type I ne sont pas strictement identiques, il est difficile de tirer des conclusions définitives.

Nous avons aussi montré que l'intérêt principal de la méthode `ALTree` est de permettre de faire des hypothèses concernant le site de susceptibilité à la maladie. Par exemple, lorsque plusieurs sites en interaction sont impliqués dans le déterminisme de la maladie, elle permet une bien meilleure localisation des sites de susceptibilité qu'une analyse locus par locus.

3.3.3 Article 2

On the use of haplotype phylogeny to detect disease susceptibility loci (Bardel *et al.*, 2005a)
Claire Bardel, Vincent Danjean, Jean-Pierre Hugot, Pierre Darlu et Emmanuelle Génin. *BMC Genetics* 2005, **6** : 24.

[6] environ 2 jours par réplicat

BMC Genetics

BioMed Central

Methodology article

Open Access

On the use of haplotype phylogeny to detect disease susceptibility loci

Claire Bardel*[1], Vincent Danjean[2], Jean-Pierre Hugot[3,4], Pierre Darlu[1] and Emmanuelle Génin[1]

Address: [1]Unité de recherche en Génétique Épidémiologique et structure des populations humaines, INSERM U535, Villejuif, France, [2]Laboratoire Bordelais de Recherche en Informatique, UMR 5800, Bordeaux, France, [3]Programme Avenir, INSERM U458, hôpital Robert Debré, AP-HP, Paris, France and [4]Fondation Jean Dausset, Paris, France

Email: Claire Bardel* - bardel@vjf.inserm.fr; Vincent Danjean - danjean@labri.fr; Jean-Pierre Hugot - jean-pierre.hugot@rdb.ap-hop-paris.fr; Pierre Darlu - darlu@vjf.inserm.fr; Emmanuelle Génin - genin@vjf.inserm.fr

* Corresponding author

Published: 18 May 2005

BMC Genetics 2005, **6**:24 doi:10.1186/1471-2156-6-24

Received: 21 January 2005
Accepted: 18 May 2005

This article is available from: http://www.biomedcentral.com/1471-2156/6/24

© 2005 Bardel et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The cladistic approach proposed by Templeton has been presented as promising for the study of the genetic factors involved in common diseases. This approach allows the joint study of multiple markers within a gene by considering haplotypes and grouping them in nested clades. The idea is to search for clades with an excess of cases as compared to the whole sample and to identify the mutations defining these clades as potential candidate disease susceptibility sites. However, the performance of this approach for the study of the genetic factors involved in complex diseases has never been studied.

Results: In this paper, we propose a new method to perform such a cladistic analysis and we estimate its power through simulations. We show that under models where the susceptibility to the disease is caused by a single genetic variant, the cladistic test is neither really more powerful to detect an association nor really more efficient to localize the susceptibility site than an individual SNP testing. However, when two interacting sites are responsible for the disease, the cladistic analysis greatly improves the probability to find the two susceptibility sites. The impact of the linkage disequilibrium and of the tree characteristics on the efficiency of the cladistic analysis are also discussed. An application on a real data set concerning the CARD15 gene and Crohn disease shows that the method can successfully identify the three variant sites that are involved in the disease susceptibility.

Conclusion: The use of phylogenies to group haplotypes is especially interesting to pinpoint the sites that are likely to be involved in disease susceptibility among the different markers identified within a gene.

Background

With the development of molecular techniques to identify genetic polymorphisms, numerous markers and in particular Single Nucleotide Polymorphisms (SNPs) [1,2] are now available within and between genes for establishing their possible role in disease susceptibility. One recurrent question in the literature is the way these markers could be used to test for association with complex diseases [3,4].

Should one focus on one marker at a time and simply compare allele or genotype distributions for the studied markers in cases and controls or should one consider haplotypes formed by several linked markers? Haplotypic methods have been suggested to be more powerful at detecting the role of a given genomic region in disease susceptibility [4-7]. Indeed, disease susceptibility may be due to the combined effects of variants at different markers. Different methods have been proposed to test for association between markers located on the same reconstructed haplotypes and disease susceptibility [4,7-10]. In these tests, the number of haplotypes being large, the degrees of freedom to compare cases and controls are also large, decreasing the power to detect the association. Moreover, as some haplotypes would only be carried by one or two individuals, there could be statistical problems owing to small sample sizes making difficult the evaluation of their possible effect on the susceptibility to the disease.

A solution to reduce these problems is to group haplotypes in order to decrease the degrees of freedom and to increase the number of individuals in the different haplotype groups. The cladistic method as described by Templeton et al. [11] is a method to carry out such groupings of haplotypes and to determine which haplotype or group of haplotypes is likely to be responsible for the phenotypic variation observed among the population. It consists in using parsimony methods to build a phylogeny of the haplotypes, and to group haplotypes according to the clades defined by the phylogenetic tree. When applied to a case/control data set, the proportions of cases and controls in different clades are compared. If a clade shows a significantly larger number of cases than the others, one will conclude, first, that there is an association between the disease and one haplotype or a group of haplotypes belonging to the clade, and, second, that the mutations defining the clade containing an excess of cases are good candidates to be functional gene polymorphisms involved in the disease susceptibility.

The cladistic analysis relies for a large part on the reconstruction of an accurate phylogenetic tree, but recombination is known to bias phylogenetic reconstruction processes [12]. The recent discovery of the haplotype block structure of the human genome [13] now provides an interesting framework for cladistic methods. Indeed, haplotype blocks define regions in which the recombination rate is quite low. Moreover, once a block is found to be associated with a disease, it will not be suitable to use classical methods based on recombination to fine map susceptibility loci. Haplotype phylogenies being based on the history of mutations may then help to identify disease susceptibility loci.

The cladistic method was first described in a series of five articles [11,14-17] and applied in different case-control studies focusing either on quantitative or qualitative data. For example, Haviland et al. used a cladistic analysis to study the association between the apolipoprotein AI-CIII-AIV and plasma lipid, lipoprotein and apolipoprotein levels [18]. Keavney et al. [19] used cladograms to localize the sites that significantly influence Angiotensin-I Converting Enzyme (ACE). Zhu et al. [20] worked on quantitative data simulated for the twelfth Genetic Analysis Workshop (GAW). Using regression analysis and analysis of variance (ANOVA), they were able to find a short region containing the simulated mutation. Wang et al. [21] used a nested linear model to analyze the association between *ApoA4* haplotypes and plasma lipid level in the cladistic framework. Among the studies that consider discrete data is the one by Kittles et al. [22] who tested the association between haplotype variations on the Y chromosome and alcohol dependence and related personality traits. Lobos and Todd applied the same cladistic method to four RFLPs in the tyroxine hydroxylase gene [23], and to five SNPs in the dopamine receptor D2 gene [24]. Darlu and Génin [25] applied the method to the GAW 12 simulated data and were able to identify some of the disease susceptibility sites. For a more detailed review on the subject, see [26].

More recently Seltman et al. [27] developed a method that can be applied on family-based data. They proposed a test, the ET-TDT, that combines the transmission disequilibrium test (TDT) and the use of unrooted phylogenetic trees to group haplotypes. They further extended their method to the analysis of qualitative and quantitative data from case-parent trios or population-based studies [28]. Finally, an alternative method where phylogenetic trees are reconstructed by a distance method was proposed by Durrant et al. [29] and implemented in the software CLADHC.

From these different examples, cladistic methods appear to be innovative methods to detect an association between a candidate gene and a disease and to localize disease susceptibility sites. However, the validity and the power of these methods to detect association have only been tested in a very limited number of situations (fixed pattern of linkage disequilibrium, fixed phylogenetic tree) [27,29]. In this paper, we develop a new cladistic test (CT), based on the analysis of a rooted tree, which allows to make hypotheses about functional polymorphisms that are involved in the susceptibility to the disease. Through simulations on ten different tree topologies, we evaluate the conditions in which this cladistic method is efficient and compare it to other methods. Finally, we apply this method to a real data set concerning Crohn disease and the CARD15 gene [30].

Table 1: Power to detect an association, one susceptibility site simulated and included in the analysis

Tree	CT	CLADHC	HT	SbST
1	96.4 *(3.3)*	88.9 *(3.3)*	91.1 *(5.1)*	99.3*(3.4)*
2	99.0 *(4.4)*	95.7 *(4.1)*	91.3 *(6.0)*	99.5*(5.3)*
3	86.0 *(2.5)*	91.1 *(4.7)*	91.2 *(4.6)*	99.3*(4.4)*
4	96.3 *(3.8)*	95.6 *(2.9)*	91.5 *(4.2)*	99.3*(4.6)*
5	88.8 *(4.1)*	95.3 *(3.3)*	90.4 *(4.7)*	99.1*(4.6)*
6	84.5 *(3.5)*	88.8 *(2.7)*	90.7 *(4.4)*	99.5*(4.4)*
7	97.4 *(4.0)*	96.3 *(3.7)*	91.9 *(4.8)*	99.2*(4.3)*
8	99.7 *(3.2)*	98.5 *(2.4)*	91.2 *(4.9)*	99.4 *(4.7)*
9	88.1 *(3.5)*	94.0 *(4.4)*	91.3 *(4.7)*	99.4*(4.3)*
10	92.5 *(2.6)*	93.0 *(2.9)*	91.3 *(4.7)*	99.5*(3.6)*
Average	92.9 *(3.5)*	93.7 *(3.4)*	91.2 *(4.8)*	99.35*(4.4)*
Std dev.[a]	5.63 *(0.62)*	3.24 *(0.76)*	0.41 *(0.49)*	0.13 *(0.54)*

Sample size: 100 cases and 100 controls, 1000 replicates, penetrance vector: [0.03;0.06;0.3], frequency of the susceptibility allele: 0.205 ± 0.005. The corresponding type I errors are indicated in italics within brackets. For each tree, the method giving the best result is underlined. CT: Cladistic Test, CLADHC: test developed by Durrant et al. [29], HT: Haplotypic Test, SbST: Site by Site Test.
[a] Standard deviation

Table 2: Power to detect an association, one susceptibility site simulated and removed from the analysis

Tree	CT	CLADHC	HT	SbST
1	94.8 *(3.8)*	89.1 *(3.5)*	91.2 *(5.6)*	89.5 *(3.9)*
2	94.0 *(3.8)*	92.7 *(4.1)*	87.6 *(5.6)*	98.5 *(5.3)*
3	79.3 *(3.2)*	88.9 *(4.7)*	86.2 *(4.7)*	94.2 *(4.1)*
4	88.7 *(2.3)*	85.7 *(3.1)*	87.8 *(4.6)*	93.6 *(4.4)*
5	85.6 *(4.3)*	94.4 *(4.1)*	90.5 *(4.5)*	94.0 *(4.1)*
6	78.6 *(3.0)*	81.6 *(3.3)*	88.0 *(4.2)*	67.0 *(4.2)*
7	92.4 *(3.8)*	90.5 *(3.7)*	87.8 *(5.2)*	92.7 *(4.0)*
8	99.7 *(3.0)*	98.6 *(2.8)*	91.2 *(4.7)*	99.6 *(4.5)*
9	91.9 *(3.7)*	89.2 *(4.5)*	91.2 *(4.8)*	89.1 *(4.1)*
10	88.8 *(2.9)*	82.8 *(3.0)*	81.3 *(4.1)*	79.1 *(3.5)*
Average	89,38 *(3.4)*	89.35 *(3.7)*	88,28 *(4.8)*	89,73 *(4.2)*
Std dev.[a]	6.71 *(0.6)*	5.2 *(0.6)*	3.06 *(0.5)*	9.81 *(0.5)*

Sample size: 100 cases and 100 controls, 1000 replicates, penetrance vector: [0.03;0.06;0.3], frequency of the susceptibility allele: 0.205 ± 0.005. The corresponding type I errors are indicated in italics within brackets. For each tree, the method giving the best result is underlined. CT: Cladistic Test, CLADHC: test developed by Durrant et al. [29], HT: Haplotypic Test, SbST: Site by Site Test.
[a] Standard deviation

Results

Association test

As detailed in the method section, four different tests are performed: a single locus allele-based test referred to as SbST (Site by Site Test), an haplotype-based test referred to as HT (Haplotypic Test), a test based on haplotype phylogenies referred to as CT (Cladistic Test) and a test based on the clustering of haplotypes described by Durrant et al. [29] and referred to as CLADHC. To compare their respective power to detect an association, simulations were performed under three disease models involving one (model 1) or two (model 2 and model 3) susceptibility sites for ten different tree topologies.

Results for model 1 are presented in Table 1 and 2. When the susceptibility site is included in the studied SNPs (Table 1), SbST is the most powerful test for all trees but one (tree 8) (between 99.1% and 99.5% depending on the tree, 99.35% on average), while the other tests have a weaker power (92.9%, 93.7%, and 91.2%, on average, for CT, HT, and CLADHC respectively). CT and CLADHC have almost the same average type I error, slightly but

Table 3: Power to detect an association for two susceptibility sites simulated under model 2

	Model 2			
Tree	CT	CLADHC	HT	SbST
1	97.8 *(3.7)*	99.3 *(3.8)*	97.7 *(5.0)*	99.1 *(6.9)*
2	95.1 *(3.6)*	97.7 *(2.9)*	97.9 *(5.6)*	95.8 *(4.3)*
3	94.2 *(4.9)*	99.4 *(4.9)*	97.9 *(5.1)*	98.6 *(4.9)*
4	97.6 *(4.6)*	97.8 *(4.6)*	97.8 *(5.5)*	98.6 *(4.6)*
5	98.5 *(5.3)*	99.5 *(4.5)*	98.9 *(5.7)*	98.2 *(7.4)*
6	98.9 *(4.4)*	97.9 *(3.9)*	98.5 *(5.7)*	96.9 *(6.6)*
7	93.0 *(5.1)*	93.5 *(4.6)*	96.9 *(5.6)*	95.5 *(5.2)*
8	95.2 *(6.1)*	99.0 *(5.3)*	98.2 *(6.2)*	95.4 *(4.2)*
9	96.7 *(3.6)*	97.3 *(3.1)*	97.6 *(4.1)*	97.7 *(4.0)*
10	99.9 *(4.5)*	99.5 *(4.0)*	98.2 *(60.)*	99.8 *(5.8)*
Average	96.7 *(4.6)*	98.06 *(4.2)*	97.96 *(5.45)*	97.56 *(5.4)*
Std dev.[a]	2.24 *(0.8)*	1.83 *(0.8)*	0.54 *(0.6)*	1.58 *(1.2)*

Sample size: 200 cases and 200 controls, 1000 replicates, penetrance: 0.9 for the homozygotes carrying the two susceptibility alleles, 0.3 for the other genotypes. The corresponding type I errors are indicated in italics within brackets. For each tree, the method giving the best result is underlined. CT: Cladistic Test, CLADHC: test developed by Durrant et al. [29], HT: Haplotypic Test, SbST: Site by Site Test

[a] Standard deviation

significantly lower than the type I error of HT and SbST. In these simulations, as the type I errors cannot be fixed a priori, powers may be difficult to compare. However, if the type I errors of CT and CLADHC were increased, their power would also be increased and, consequently, the difference in power with HT would be even more pronounced and the difference in power with SbST would be more reduced. In the trees showing comparable type one error (Trees 1, 5, 7), the power of SbST is clearly higher than the power of CT. When the susceptibility site is removed (Table 2), the performances of the four tests are, on average, similar (about 89%), CT being the most powerful test for four trees, SbST for four other trees, HT and CLADHC for one tree each. However, the type I errors being slightly different, by adjusting them one expects that CT and CLADHC could be on average, slightly more powerful than HT and SbST.

As expected, the power is higher when the susceptibility site is included than when it is removed. In this latter case, the power depends on the degree of linkage disequilibrium (LD) between the susceptibility site and the other sites as expressed by the maximum value of LD, referred to as LD_{max}. Moreover, since the pattern of LD can be different from one tree to another, one expects more differences in power between trees when the susceptibility site is removed. Indeed the standard deviation of the power is larger in this case. This standard deviation is also particularly large for SbST (σ = 9.81). This test, being an allele-based test, is more influenced by LD than the three others tests that are haplotype-based tests, as proved by the high positive correlation observed between the power of SbST to detect association and the LD_{max} (Spearman rank correlation ρ equals 0.851, p-value < 0.01). For CT and CLADHC, Spearman rank correlations are not significant ($p > 0.1$) confirming that LD has a lesser impact on these two tests. Therefore, the high standard deviations observed for these two tests are probably more related to the phylogeny of haplotypes than to the linkage disequilibrium.

The results for model 2 and model 3, differing by the values of the penetrance vector, are presented in Table 3 and in Table 4 respectively. There is no real difference in power between the four tests for model 2 (about 97%). For model 3, the power of the four tests is approximately half the power for model 2 owing to the reduced value of the penetrance for the genotype associated to the highest risk in this model. In this case, SbST and CT are, on average, the most and the less powerful test respectively

Localization of the susceptibility site(s)

The efficiency of SbST, CT and CLADHC to localize the susceptibility site(s) when it is included among the SNPs is presented in Figure 1 and 2. The efficiency of CT is plotted against the efficiency of SbST or CLADHC. To correctly interpret these figures, it has to be noted that the criteria used to compare the methods are not strictly identical. Indeed, the identified SNP is always unambiguously detected by the SbST method, since this is the site giving the most significant chi-square and two chi-squares are never exactly identical. On the other hand, the

Table 4: Power to detect an association for two susceptibility sites simulated under model 3

	Model 3			
Tree	CT	CLADHC	HT	SbST
1	40.7 *(4.7)*	54.0 *(4.5)*	42.0 *(5.9)*	47.5 *(4.7)*
2	31.7 *(3.4)*	32.7 *(3.9)*	43.0 *(5.9)*	45.8 *(5.3)*
3	26.8 *(4.2)*	47.9 *(4.4)*	40.7 *(6.4)*	45.2 *(3.5)*
4	34.8 *(4.8)*	39.2 *(4.7)*	38.9 *(7.0)*	47.0 *(4.4)*
5	41.5 *(5.6)*	49.7 *(4.7)*	44.2 *(6.6)*	37.2 *(4.6)*
6	42.0 *(3.6)*	34.8 *(4.7)*	42.4 *(5.6)*	41.3 *(5.3)*
7	28.6 *(5.2)*	32.1 *(4.7)*	41.6 *(6.4)*	39.9 *(4.3)*
8	33.2 *(4.3)*	47.2 *(4.0)*	42.9 *(6.1)*	37.7 *(4.0)*
9	33.5 *(4.4)*	39.1 *(4.1)*	41.2 *(7.0)*	48.5 *(5.3)*
10	65.6 *(6.6)*	48.1 *(4.1)*	42.1 *(7.8)*	56.9 *(4.2)*
Average	37.84 *(4.7)*	42.48 *(4.4)*	41.9 *(6.5)*	44.7 *(4.6)*
Std dev.[a]	11.07 *(0.9)*	7.84 *(0.3)*	1.45 *(0.7)*	5.94 *(0.6)*

Sample size: 200 cases and 200 controls, 1000 replicates, penetrance: 0.6 for the homozygotes carrying the two susceptibility alleles, 0.3 for the other genotypes. The corresponding type I errors are indicated in italics within brackets. For each tree, the method giving the best result is underlined. CT: Cladistic Test, CLADHC: test developed by Durrant et al. [29], HT: Haplotypic Test, SbST: Site by Site Test
[a] Standard deviation

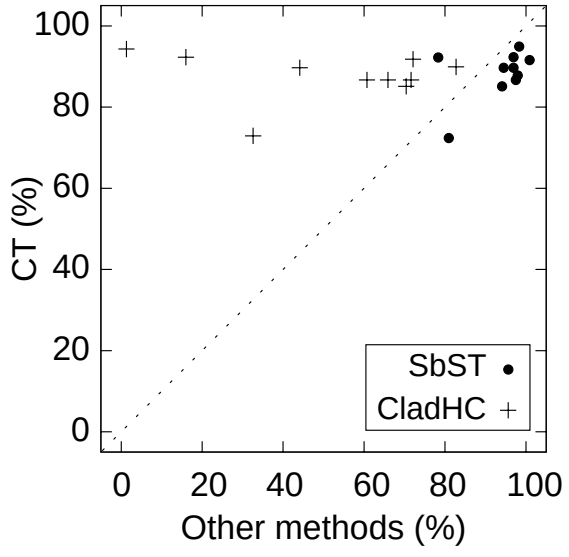


Figure 1
Comparison of the efficiency to localize the susceptibility site between CT, and SbST or CLADHC. The efficiency to localize the susceptibility site is measured by the percentage of replicates in which the simulated susceptibility site is either uniquely detected or detected among other putative sites. black circles: comparison of CT and SbST, crosses: comparison of CT and CLADHC. Each symbol corresponds to one of the ten tree topologies.

susceptibility site may not be the only one detected by CT because two or more sites could have the same values of the criteria (V_i, see the Methods section). However, in our simulations, the susceptibility site is the only one detected in from 80% (for 1 tree) to 100% (for 5 trees) of the replicates. When the susceptibility site is not found alone, it is almost always found with only one other site, rarely with two other sites (<0.5% of the replicates).

For model 1 (Figure 1), CT is slightly less efficient than SbST (for 9 trees out of 10) and more efficient than CLADHC (for all 10 trees). For SbST, we note a high negative correlation between the efficiency to localize the susceptibility site and the LD_{max} between the susceptibility site and the other sites (Spearman rank correlation ρ equal -0.88 with a p-value of 0.008). It shows that the presence of a site in relatively high LD with the susceptibility site disturbs the localization process, increasing the uncertainty of the susceptibility site's identification. Interestingly, there is no such statistically significant correlation with CT and CLADHC ($p > 0.08$), due to the fact that these last two tests use both the information on the whole haplotypes and on their history through the tree reconstruction. The presence of sites in high LD with the susceptibility site does not prevent these two methods from localizing correctly the susceptibility site.

For model 2 and model 3, (Figure 2A and 2B respectively), CT is more or at least as efficient as SbST to localize the two sites, whatever the penetrance model. For the detection of at least one of the two sites, we can observe that,

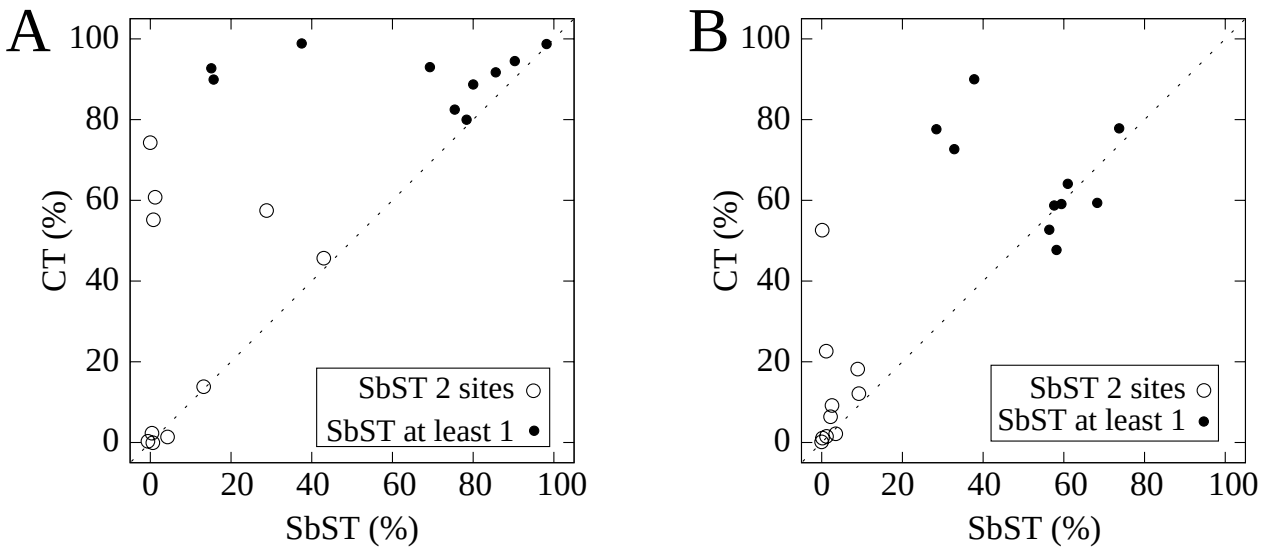


Figure 2
Comparison of the efficiency to localize the two susceptibility sites between CT and SbST. The efficiency to localize the two susceptibility sites is measured by the percentage of replicates in which the two sites are detected (black circles) or at least one of the two sites (white circles) Each symbol corresponds to one of the ten tree topologies. (A): penetrance 0.9 for homozygotes carrying the two susceptibility alleles and 0.3 for all other genotypes; (B): penetrance 0.6 for homozygotes carrying the two susceptibility alleles and 0.3 for all other genotypes.

whatever the tree, CT is always more efficient than SbST under model 2, whereas this is only true for 6 out of the 10 trees under model 3.

Application to CARD15/NOD2 polymorphisms and Crohn disease

The reconstruction of the phylogenetic tree of the 33 haplotypes leads to a very high number of equally parsimonious trees. One thousand of them were retained and analyzed. In table 5, for the two rooting methods mfH and consH, the SNPs are ranked from the least to the most significant (based on the V_i parameter, see the Methods section). Association is detected with CT ($p_{mfH} = 2 \times 10^{-5}$, $p_{consH} = 3 \times 10^{-5}$) and SNP 8 is selected as potential susceptibility site. If haplotypes carrying the susceptibility allele at SNP 8 are removed, an association can still be detected ($p_{mfH} = 4 \times 10^{-5}$, $p_{consH} = 3 \times 10^{-5}$) and SNP 12 is identified as potential susceptibility site. When removing haplotypes with the susceptibility allele at SNP 12, there is still a remaining association ($p_{mfH} = 8 \times 10^{-5}$, $p_{consH} = 3 \times 10^{-5}$) and depending on the rooting method either SNP 13 (with consH) or SNP 9 (with mfH) are retained. When haplotypes with the susceptibility allele at SNP 13 are removed, the association test is not significant ($p_{mfH} = 2.9 \times 10^{-1}$, $p_{consH} = 1.1 \times 10^{-1}$). When haplotypes with the susceptibility allele at SNP 9 are removed, the association test is significant ($p_{mfH} = 2 \times 10^{-5}$, $p_{consH} = 5 \times 10^{-5}$) and the identified site is SNP 13 for both rooting method. In the end, the combination of SNP 8, SNP 12 and SNP 13 is the most parsimonious combination that can explain the susceptibility to the disease. Moreover, we can note that these 3 SNPs are those showing the highest V_i on the whole data set. So, we finally have identified three SNPs involved in the susceptibility to Crohn's disease.

Discussion

In this paper, we have presented a method to perform phylogeny-based nested haplotype analysis of case/control data and introduced the cladistic test (CT). We have shown by simulations that the CT test is approximately as powerful to detect an association as a test that compares haplotype distributions in cases and controls with no prior grouping. However, a major advantage of CT towards other haplotypic tests is the possibility of making inferences on potential susceptibility sites. Indeed the method of parsimony used to build the phylogeny allows the reconstruction of character state changes on the tree and thus the identification of the most likely susceptibility sites. Finding the polymorphisms responsible for the susceptibility to a disease is an important issue in genetic

Table 5: Cladistic analysis of Crohn's data 1000 equiparsimonious trees are studied. SNPs are ranked from the least to the most significant. In bold, the presumed susceptibility sites previously identified by Hugot et al. [30].

	least significant				-------->	most significant		
mfH[a]	SNP 5	SNP 9	SNP 6	SNP 7		**SNP 13**	**SNP 12**	**SNP8**
consH[b]	SNP 9	SNP 5	SNP 6	SNP 7		**SNP 13**	**SNP 12**	**SNP8**

[a] Rooting on the most frequent haplotype
[b] Rooting on a consensus sequence

epidemiology and in the past few years, different methods have been proposed that use linkage disequilibrium to fine-map candidate genes and to localize disease mutations ([31,32] and [33] for example) These methods are based on the fact that a disease variant arises on a particular haplotype, the haplotype and the disease thus being initially associated. Then, over time, this association is broken by recombination and only a small region of the ancestral haplotype is preserved around the disease variant. In such models, recombination is the event of interest. Other parameters such as mutations can be integrated in the models, but they are considered as nuisance parameters. The cladistic test developed here is different in the sense that it is based on the reconstruction of the evolutionary history of the mutations that occur to form the haplotypes. Contrary to the other fine-mapping methods, the event of interest is mutation, and recombination is a nuisance parameter. Through the simulation study, we have shown that this test is a powerful tool to suggest etiological SNPs. This is also illustrated here on Crohn disease since the cladistic test was able to correctly identify the 3 sites within the CARD15 gene that are already reported to be involved in the disease susceptibility [30].

The cladistic test requires the building of a phylogeny of the different haplotypes present in the sample. Different methods are available to build haplotype phylogenies: distance methods, parsimony methods and maximum likelihood (ML) methods. A parsimony method was chosen here because it allows the reconstruction of apomorphies that can be used to define potential sites involved in the disease susceptibility and because it runs faster than ML methods, thus allowing a simulation study. We have compared CT with another clustering method, CLADHC, a distance-based method that also differs from CT in the statistical analysis (CLADHC performs regression analysis, CT is based on chi-square tests). We have shown that the power of these two tests to detect association is quite similar, CT being slightly more powerful when one site is simulated, and CLADHC when two sites are simulated. For the localization of the susceptibility site, the efficiency of the two methods is difficult to compare because CT directly localizes the site on short sequences while CLADHC localizes a window containing the susceptibility site on longer sequences. The two test should thus be used in different situations. Nevertheless CT appears to be more efficient in our simulation conditions

In the phylogeny reconstruction, one is faced to different problems. The first one is the choice of the ancestral sequence to root the phylogeny. With real data, the ancestral sequence is usually not known, unless at least two outgroups can be postulated (in the parsimony context). In the absence of ancestral sequences, hypotheses have to be made, and they are numerous, like taking the most frequent haplotype, but this haplotype could be different from one population to an other or from one sampling to another, and this most frequent haplotype may not have a significantly different frequency from the next most frequents. Or one can take the consensus haplotype, but this sequence may not be observed in the sample, or not be among the most frequent sequences. In this study, our strategy was to consider the uncertainty of the ancestral sequence as a nuisance parameter, since our goal was not primarily to focus on this problem and because we must be in the same situation for all the performed simulations. Extensive further simulations will be needed to assess the properties of the different rooting methods. However, several ancestral hypotheses should be considered, discussed, and tested on real data, as we have done for the Crohn data. Second, the presence of recombinations would restrict the use of phylogenetic reconstruction. Indeed, phylogenetic reconstruction is known to be biased by recombinations [12] in a manner depending on how and when the recombination events occur in the tree. This context led us to simulate the history of mutations along the haplotypes without introducing recombinations between haplotypes. However, because in our method the reconstructed tree has not to represent the exact history of the haplotypes, but only gives a way to cluster them in an appropriate way, we believe that our method can tolerate some low degrees of recombination, as it is probably the case for the Crohn data. Further simulations would however be required to better assess this point. Third, the choice of the way characters are optimized on the tree is another important point in cladistic analysis. Two classical options can be used to optimize the character state changes along the tree: deltran, (delayed transformation)

that favors convergences or acctran(accelerated transformation) that favors reversions. In this paper, only the results performed with the deltran option are presented. Simulations (one susceptibility site) have also been performed with the acctran option. For the association detection, the results are very similar to those obtained with deltran showing that this parameter does not influence the power of the CT method to detect association in our simulation conditions. For the localization of the susceptibility site, the results are different. The method is more efficient with the deltran option than with the acctran option for eight trees out of ten, it is less efficient for one tree and the results are equivalent for the last tree. It is partly due to to the uncertainties of the character *S* and to the frequency of the disease susceptibility allele. When the frequency of the susceptibility allele is low, as in our simulations, and when *S* is coded "?" in some nodes of the tree, one expects the change of *S* from 0 to 1 to occur later in the tree, and the delayed option (deltran) will then be more appropriate. Conversely, when the frequency of the disease is large, the change of *S* from 0 to 1 has more chances to occur earlier in the tree since the *S* mutation can then include more cases, and the accelerated option (acctran) will be more appropriate. However, when analyzing a real data set, it is advisable to compare the results obtained with both options.

The variability of the results obtained with the cladistic method is due for a minor part to the LD between the susceptibility site(s) and the other sites and for the major part to parameters specific to the tree. Among these parameters, the position of the mutation in the tree (near the root or the leaves, sample size of the clades near the mutations...) has already been reported to have an influence on the power of cladistic methods to detect association [27]. Apart from the position of the mutation, the presence of sites that co-mutate with the susceptibility site, the degree of multiple mutations of the susceptibility site in the tree, and the existence of one or more reversions for the susceptibility site influence the power of the cladistic test to detect an association, and they have an even greater impact on the efficiency of the cladistic test to localize the susceptibility site. As all these factors are not independent, the evaluation of their own effects remains problematic.

The four methods compared here (CT, HT, SbST and CLADHC) suppose that haplotypic data are available, which is not usually the case. The most likely haplotypes for an individual have to be reconstructed using numerical methods (for a review of these methods, see [34,35]). The availability of familial data may help to determine the phase of the markers and then, to obtain the most likely haplotypes by using the genetic information of the two parents. However, even when these familial data are not available, the numerical methods have been shown to be very efficient to reconstruct haplotypes, except for rare haplotypes that may either be missed or falsely created. In this work, since we choose to focus on the comparison between several methods, the problem of haplotype inference is ignored. Moreover, as the four studied methods are all based on already known or inferred haplotypes carried by the individuals, the uncertainty due to the haplotype inference should not play differently from one method to another. Further extension of the CT method will take this uncertainty into account by using, for example, likelihood-based tests.

Conclusion

In conclusion, as proved by our simulations and the Crohn example, the use of phylogenies to group haplotypes by clades and the nested analysis of case-control data on the basis of these clades turns out to be an interesting and complementary strategy for the genetic study of multifactorial diseases, especially to pinpoint the sites that are likely to be involved in disease susceptibility among the different markers identified within a gene.

Methods

Principle of the method

The material on which the method is based is a sample of haplotypes composed of a combination of SNPs, each haplotype being labeled either as control or as affected depending on the phenotype of the individual. After building the phylogenetic tree of these different haplotypes, a series of nested homogeneity tests are performed to detect differences in the distribution of cases and controls in the different clades. Once an association is detected, a new character is defined according to the proportion of cases carrying an haplotype. Then, it is optimized on the tree and the sites that significantly co-mutate with this new character are putative susceptibility sites for the disease.

Simulation study

To test the power of this method to detect an association and its efficiency to precisely localize the susceptibility site, we performed the following steps : i) simulation of the haplotypes and choice of the susceptibility site(s) (one or two susceptibility sites are considered) ; ii) attribution of a status (case or control) to these haplotypes taking into account the genetic model of the disease; iii) reconstruction of the phylogeny of haplotypes and inference of the character state changes along the branches of the tree; iv) association test: nested analysis of the case/control ratio in the clades; v) localization of the susceptibility site using the equiparsimonious (ie, all the trees that have the minimum number of character state changes) ; vi) estimation, by simulations, of the power to detect the association between haplotypes and the disease and comparison with

other methods; vii) estimation of the efficiency to localize susceptibility sites and comparison with other methods.

Haplotype simulation

24 different haplotypes are simulated using the TREEVOLVE software (version 1.32) [36]. An ancestral haplotype sequence of 20 sites is randomly produced, a guide tree is generated using the coalescent model without recombination and the evolution of this ancestral sequence along the tree is simulated to obtain a sample of 24 different haplotypes. Since TREEVOLVE can only simulate the evolution of DNA sequences (4 nucleotides) and not the evolution of bi-allelic markers, we set the frequencies for bases *A* and *T* to very low values (*A* = *T* = 0.001 and *G* = *C* = 0.498) in order to obtain an ancestral sequence composed only of *G* and *C*, as an equivalent of SNPs coded by 0/1 character states. The mutation rate is set to 2×10^{-6}, and the substitution rates are chosen so that only G-C transversions could occur: *A* - *C* = 0.001, *A* - *G* = 0.001, *A* - *T* = 0.001, *C* - *T* = 0.001, *G* - *T* = 0.001, *G* - *C* = 0.995. No heterogeneity in the substitution rate among sites is specified. The population size was set to 100,000 and its growth was stationary with no subdivision. Once the haplotypes are obtained by simulation, the susceptibility sites are chosen among the 20 sites. For the simulations with only one susceptibility site, the program picks up the first site in the sequence having a minor allele frequency (p) in the sample of haplotypes within the range $F < p < F + 0.01$ where F is a predefined allele frequency (in our simulations, $F = 0.20$). For two simulated sites, the program picks up the first two simulated sites verifying a user-defined constraint. In our simulations, the two sites are chosen such that $f(A_1B_1) = f(A_1B_2) = f(A_2B_1) = f(A_2B_2) = 0.25$, A_1 and A_2 being the two alleles at the first site and B_1 and B_2, the two alleles at the second site. When no site is found satisfying these conditions, a new set of haplotypes is simulated until the required conditions are obtained.

This entire procedure is repeated ten times to generate ten different data sets. The pairwise linkage disequilibrium (r^2 measure, mean values for the 1000 replicates) and the tree topologies corresponding to these ten data sets are provided as Additional file 1 and 2.

Attribution of disease status

To generate the genotypes of each individual, pairs of haplotypes are sampled at random with replacement and the disease status (case or control) is determined by his genotype at the susceptibility site(s) and by the penetrances. In our simulations, three disease models are studied, involving either one susceptibility site (model 1) or two interacting susceptibility sites (model 2 and model 3). The first model is the disease model described in Bourgain et al. [37] that roughly corresponds to the effect of *ApoE4* in Alzheimer's disease [38]. We set the penetrance of the three genotypes *DD*, *Dd* and *dd* to respectively $P_{DD} = 0.03$, $P_{Dd} = 0.06$, $P_{dd} = 0.30$, d being the disease susceptibility allele. For model 2 and model 3, an at-risk haplotype formed by the combination of the alleles A_2 at the first locus and B_2 at the second locus is defined. The penetrance associated to the genotype $A_2B_2A_2B_2$ is set to 0.9 for model 2 and to 0.6 for model 3. All the other penetrances are set to 0.3 for both models.

The sampling process is carried on until N cases and N controls are obtained (N = 100 or 200 in our simulations, depending on the number of simulated susceptibility sites). This constitutes a replicate. As the haplotypes are sampled randomly from the set of 24 different haplotype, they are almost equi-frequent in the sample of cases and controls, and their frequencies are close to $\frac{1}{24} = 0.042$.

Phylogeny of haplotypes

All the different haplotypes are used to build a phylogenetic tree by a parsimony method which consists in choosing the tree that minimizes the number of mutation events [39]. The PAUP software (version 4.0b10 [40]) is used to reconstruct the tree. A heuristic search is performed, using stepwise addition and TBR (tree bisection and reconnection). All sites are equally weighted, and changes from allele 0 to allele 1 and from allele 1 to allele 0 are equally allowed for all sites, including the susceptibility site (unordered option). Among the list of all parsimonious trees, only the first one obtained is used for the association test, for computation time reasons and because we found that identical results are obtained for more than 98% of the equiparsimonious trees (data not shown). For the localization of susceptibility sites, all the equiparsimonious trees (or up to 100 different trees if there are too many equiparsimonious trees) are analyzed. The trees are rooted using the real ancestor obtained with TREEVOLVE. The character states are inferred along the branches by using the classical option deltran (delayed transformation), but acctran (accelerated transformation) was also investigated.

Association test

Starting from the root of the tree, series of nested homogeneity tests comparing the number of cases and controls in different clades are performed. The principle of the method is explained in Figure 3. Briefly, at each level of the tree, homogeneity in the distribution of cases and controls is tested among all the n clades defined at this level. If the test is significant, an association is detected and the analysis ends. If the test is not significant, one homogeneity test is performed between all the sub-clades descending from the n clades.

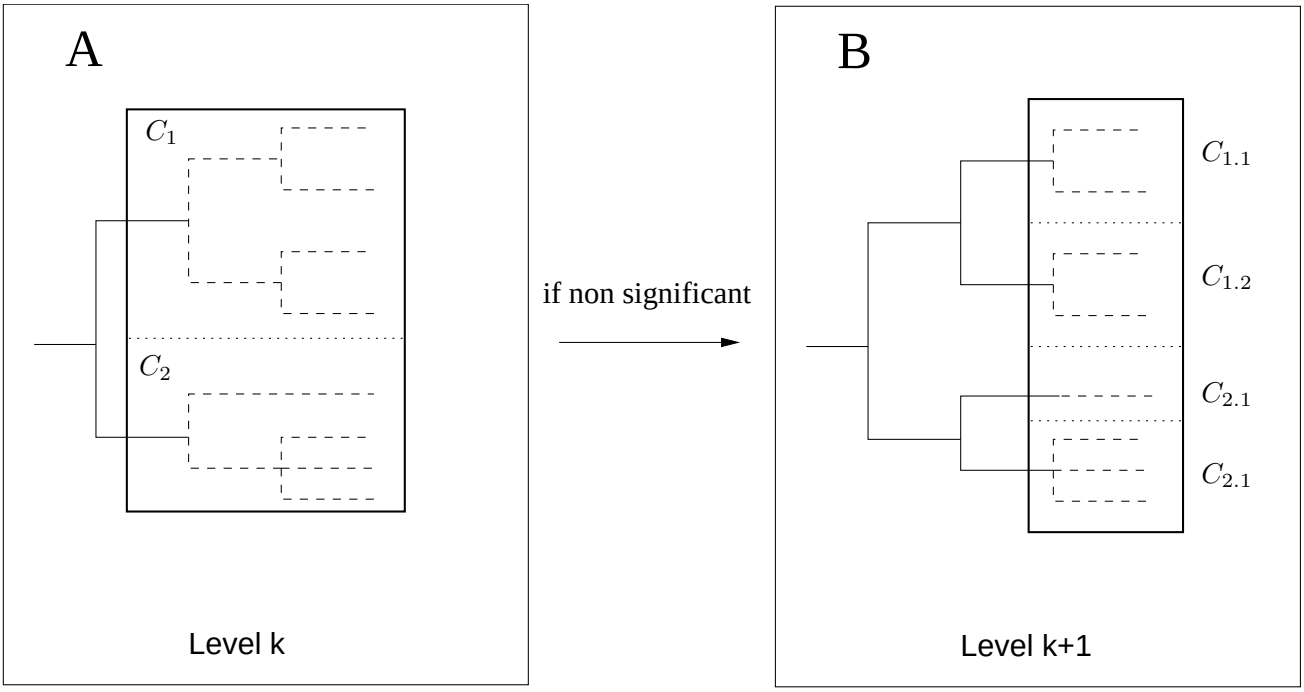


Figure 3
Description of the nested clade analysis. (A) shows the homogeneity test performed at level k (between clades C_1 and C_2). If it is not significant (B), a test will be performed at the following level (k+1), between all the sub-clades descending from clades C_1 and C_2, i.e between clades $C_{1.1}$, $C_{1.2}$, $C_{2.1}$ and $C_{2.2}$ (3 degree of freedom). If it is significant the analysis ends because an association is detected.

P-values of the homogeneity tests are calculated by Pearson chi-square tests except when only two clades are compared in which case a Fisher exact test is performed, and except when more than two clades are tested and the sample size is less than four in a category. In this latter case p-values are estimated by permutation tests.

The influence of the LD on the power of the different tests is assessed by using the r^2 measure of LD.

Localization of the susceptibility sites
For each replicate, several equiparsimonious trees may exist. We will analyze the T first equiparsimonious given by PAUP, T being limited to 100 for computation time reasons. For a tree t, a new character S is allocated to each haplotype h. The state of S is "0", "1" or "?" depending on the proportion (p_h) of cases carrying the haplotype h compared to the proportion p_0 of cases in the whole sample.

if $p_h < p_0 - \sqrt{\frac{p_h \times (1-p_h)}{n_h}}$, S is coded "0" (high number of controls);

if $p_h < p_0 - \sqrt{\frac{p_h \times (1-p_h)}{n_h}}$, S is coded "1" (high number of cases);

else, S is coded "?" (missing data).

with n_h being the number of individuals carrying the haplotype h.

This character S is optimized on the T equiparsimonious trees using PAUP and the deltran option. For each site i, let R'_{it} and R''_{it} be the observed number of times each transition (0→1 for R'_i and 1→0 for R''_i) co-mutates with a 0→1 change of the character S on tree t. Let E'_{it} and E''_{it} be the expected number of co-mutations on tree t under the hypothesis of a random distribution of the mutations on tree t and an equal probability of mutation on each branch.

$$E'_{it} = \frac{m'_{it} \times s_t}{b_t}$$

where m'_{it} (resp. m''_{it}) is the number of 0→1 (resp. 1→0) transitions of the site i on tree t;

s_t is the number of 0→1 transition of the character S on tree t;

b_t is the number of branches of tree t.

Then, for each site i on tree t, we measure the correlated evolution of the site i and the character S, by defining V'_{it} (resp. V''_{it}) as follow:

$$\begin{cases} V'_{it} = 0 & if \quad E'_{it} = 0 \\ V'_{it} = \dfrac{R'_{it} - E'_{it}}{\sqrt{E'_{it}}} & it \quad E'_{it} \neq 0 \end{cases}$$

then, for the T equiparcimonious trees, we define V'_i (resp. V''_i) as follow:

$$V'_i = \sum_{t=0}^{T} V'_{it}$$

Finally, V_i is defined as the max between V'_i and V''_i. The site or the two sites corresponding to the highest V_i are selected as putative susceptibility sites

Estimation of power and efficiency

For each of the ten samples of 24 haplotypes simulated along the ten different tree topologies, the process detailed above (attribution of disease status, phylogenetic reconstruction, association test and localization of the susceptibility site) is repeated 1000 times to obtain 1000 replicates, each of them including N cases and N controls. The power to detect the association is measured by counting the proportion of replicates in which a significant heterogeneity in the distribution of cases and controls is detected between clades. The efficiency of CT to localize the true susceptibility site is simply evaluated as the number of times, among the 1000 simulations, the true susceptibility site is among the detected sites. When two susceptibility sites are simulated, we evaluate the number of times the two sites are detected and the number of times at least one of the two sites is detected.

The test using the cladistic method (CT) is compared to three other association tests: i) a haplotypic test (HT) that compares the haplotype distributions in cases and controls: if h different types of haplotypes are observed, a chi-square test with h - 1 degree of freedom (df) is performed. As with the CT test, when sample sizes are small, a permutation procedure is used to perform the chi-square test. This HT test is only used to test for association and not to localize the susceptibility sites. ii) An allelic test, referred to as Site by Site Test (SbST), that compares the allele distributions in cases and controls at each site by a chi-square: if the number of sites is s, then s chi-squares, each with one df, are performed. The site or the two sites that give the most significant results are proposed as susceptibility sites. iii) A test based on CLADHC, the program developed by Durrant et al. [29]. In this test, the data set is analyzed by overlapping windows sliding along the haplotypes. In each window, a tree is reconstructed using a distance method, and a regression analysis is performed at each level of the tree (nested analysis). The level of the tree that maximizes the evidence of disease marker association in the likelihood-ratio test is selected. For each window, the program also provides a significant threshold calculated using the Bonferroni correction. In our simulations, a window size of 6 is selected. We consider the association as detected when at least one of the windows is significant. Concerning the localization of the susceptibility site, CLADHC can only find a window containing the susceptibility site, but not the susceptibility site itself. Therefore, we only use CLADHC in the case of one susceptibility site and we consider the localization as correct when the window with the maximum value of the statistic contains the susceptibility site.

For CT, SbST and CLADHC, we are confronted to a problem of multiple testing. When a real data set is analyzed, a permutation procedure can be used to estimate p-values of the association tests that are corrected for multiple testing. This procedure is very time-consuming and cannot be used on the simulations we have performed here. As an example, the 100,000 repetitions for the Crohn data run in about 24 hours on a Pentium III, 930 MHz, 512 Mo of RAM. Therefore, for SbST and CT, we estimate the nominal type I error required to obtain a global type-one error of 5% by a simulation under the null hypothesis of no association between the studied sites and the disease (5000 replicates). We found that for the different simulation conditions studied here, these nominal errors are equal to 1% for CT and varies from 0.2 to 0.4% for SbST. For the analysis with CLADHC, the program provides a significant threshold calculated using Bonferroni correction. As the Bonferroni correction is very conservative, to obtain similar type I errors with CLADHC as with CT, nominal type I errors were set to values varying from 15 to 17% for simulations with one susceptibility site and from 20 to 25% for simulations with two susceptibility sites.

Analysis of real data on Crohn disease

The data set is described in a paper of Hugot et al. (2001) [30] and includes 232 nuclear families with two affected children and their parents genotyped for 13 SNPs in the CARD15/NOD2 gene. Haplotypes were reconstructed in these families using GENEHUNTER 2.0b[41]. In each family, one affected child is selected at random to create the case sample (his two haplotypes are labeled cases). The control sample is formed with the parental haplotypes non transmitted to the selected children. The phylogenetic reconstruction assumes that no recombination has occurred between the SNPs. To determine if this was true for the different SNPs tested here or at least for a subset of them, linkage disequilibrium (LD) is estimated between all the SNPs. From the pairwise LD, a group of 7 SNPs (5, 6, 7, 8, 12, 9 and 13) can be identified as forming a block in strong LD among which no or very few recombinations are likely to have occurred (see Additional file 3). Interestingly, this same block of 7 SNPs was described by Vermeire et al. [42] in the CARD15/NOD2 gene. Only these 7 SNPs that form 33 different haplotypes are used in the cladistic analysis. The phylogenetic trees are reconstructed using the same parameters as in the simulations. However here, contrary to the simulations, the ancestral sequence being not known, two different analysis are performed using as ancestral sequence either the most frequently observed haplotype in our sample (mfH) or a consensus sequence formed by the combination of the most frequent alleles at each SNP (consH). For the association test, p-values are estimated on the first equiparsimonious tree obtained by PAUP, using an algorithm similar to the one described in a paper of Becker and Knapp (2004) [43]. This permutation process corrects for multiple testing in a lesser conservative way than the Bonferroni correction because it takes into account the dependency between the tests. For the localization, among all the equiparsimonious trees, 1000 of them are retained and analyzed with CT.

Authors' contributions

CB perfomed the simulation analyses, contributed to the conception of the study, to the elaboration of the algorithm and wrote the manuscript. VD contributed to the elaboration of the algorithm. J-PH provided the CARD15 data. PD and EG contributed to the design and to the conception of the study and to the manuscript preparation.

Additional material

Additional File 1

Average pairwise linkage disequilibrium for the 10 simulated date sets. *The average pairwise LD (over 1000 replicated) is calculated using the r^2 measure for each of the 10 tree topologies (T1 to T10). The matrices are plotted with GOLD[44].*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2156-6-24-S1.pdf]

Additional File 2

Tree topologies of the 10 simulated date sets

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2156-6-24-S2.pdf]

Additional File 3

Pairwise linkage disequilibrium for the Crohn data. *(A) shows the pairwise linkage disequilibrium (LD) calculated with the* D' *measure. (B) shows the LD calculated with the r^2 measures. The matrices are plotted with GOLD [44]. A high degree of LD can be observed. The differences between the r^2 and the D' values are explained by the difference in the allelic frequency for the different SNPs.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2156-6-24-S3.pdf]

Acknowledgements

We wish to thank Marie-Claude Babron and two anonymous reviewers for helpful comments and Arnaud Legrand for letting us use one of his perl function. We are also grateful to Habib Zouali and Suzanne Lepage from the fondation Jean Dausset for providing us with Crohn data and to Caroline Durrant for kindly providing us with the software CLADHC. Many thanks to the runtime team from the LaBRI for letting us run a part of our simulations on their cluster.

References

1. Collins FS: **Preparing health professionals for the genetic revolution.** *Jama* 1997, **278:**1285-1286.
2. Collins FS, Brooks LD, Chakravarti A: **A DNA polymorphism discovery resource for research on human genetic variation.** *Genome Res* 1998, **8:**1229-1231.
3. Schork NJ, Fallin D, Lanchbury S: **Single nucleotide polymorphisms and the future of genetic epidemiology.** *Clin Genet* 2000, **58:**250-264.
4. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: **Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals.** *Hum Hered* 2002, **53:**79-91.
5. Akey J, Jin L, Xiong M: **Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain?** *European J of Hum Genet* 2001, **9:**291-300.
6. Douglas JA, Boehnke M, Guillanders E, Trent JM, Gruber SB: **Experimentally-derived haplotyes substantially increase the efficiency of linkage disequilibrium studies.** *Nat Genet* 2001, **28:**361-364.
7. Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ: **Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's Disease.** *Genome Res* 2001, **11:**143-151.

8. Tiret L, Amouyel P, Rakotovao R, Cambien F, Ducimetière P: **Testing for association between disease and linked marker loci: a log-linear-mode analysis.** *Am J Hum Genet* 1991, **48:**926-934.
9. Nielsen DM, Ehm MG, Zaykin DV, Weir BS: **Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype association.** *Genetics* 2004, **168:**1029-1040.
10. Zhao JH, Curtis D, Sham PC: **Model-free analysis and permutation tests for allelic associations.** *Hum Hered* 2000, **50:**133-139.
11. Templeton AR, Boerwinkle E, Sing CF: **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila.** *Genetics* 1987, **117:**343-351.
12. Schierup MH, Hein J: **Consequences of recombination on traditional phylogenetic analysis.** *Genetics* 2000, **156:**879-891.
13. Gabriel SB, Scaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyamo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296:**2225-2229.
14. Templeton AR, Sing CF, Kessling A, Humphries S: **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations.** *Genetics* 1988, **120:**1145-1154.
15. Templeton AR, Crandall KA, Sing CF: **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation.** *Genetics* 1992, **132:**619-633.
16. Templeton AR, Sing CF: **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analysis with cladogram uncertainty and recombination.** *Genetics* 1993, **134:**659-669.
17. Templeton AR: **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction en-donuclease mapping or DNA sequencing. V. analysis of case/control sampling designs: Alzheimer's disease and the Apolipoprotein E locus.** *Genetics* 1995, **140:**403-409.
18. Haviland MB, Kessling AM, Davignon J, Sing CF: **Cladistic analysis of the Apolipoprotein AI-CIII-AIV gene cluster using a healthy French Canadian sample. I. Haploid analysis.** *Ann Hum Genet* 1995, **59:**211-231.
19. Keavney B, McKenzie CA, Connell JMC, Julier C, Ratcliffe PJ, Sobel E, Lathrop M, Farrall M: **Measured haplotype analysis of the angiotensin-I converting enzyme gene.** *Hum Mol Genet* 1998, **7(11):**1745-1751.
20. Zhu X, Cooper RS, Chen G, Luke A, Eltson R: **Localization of the Q1 mutation by cladistic analysis.** *Genet Epidemiol* 2001, **21(Suppl 1):**S594-S599.
21. Wang GQ, Pietro MD, Roeder K, Heng CK, Bunker CH, Hamman RF, Kamboh MI: **Cladistic analysis of human Apolipoprotein A4 polymorphisms in relation to quantitative plasma lipid risk factors of coronary heart disease.** *Ann of Hum Genet* 2003, **67:**107-124.
22. Kittles RA, Long JC, Bergen AW, Eggert M, Virkkunen M, Linnoila M, Goldman D: **Cladistic association analysis of Y chromosome effects on alcohol dependence and related personality traits.** *Proc Natl Acad Sci* 1999, **96:**4204-4209.
23. Lobos EA, Todd RD: **Cladistic analysis of disease association with tyroxine hydroxylase : Application to manic-depressive disease and alcoholism.** *Am J Med Genet* 1997, **74:**289-295.
24. Lobos E, Todd RD: **Association analysis in an evolutionary context: cladistic analysis of the DRD2 locus to test for association with alcoholism.** *Am J Med Genet* 1998, **81:**411-419.
25. Darlu P, Génin E: **Cladistic analysis of haplotypes as an attempt to detect disease susceptibility.** *Genet Epidemiol* 2001, **21(Suppl 1):**S602-S607.
26. Heng CK, Low PS: **Cladistic analysis: its application in association studies of complex diseases.** *Ann Acad Med Singapore* 2000, **29:**313-321.
27. Seltman H, Roeder K, Devlin B: **Transmission/Disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes.** *Am J Hum Genet* 2001, **68:**1250-1263.
28. Seltman H, Roeder K, Devlin B: **Evolutionary-based association using haplotype data.** *Genet Epidemiol* 2003, **25:**48-58.
29. Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP: **Linkage disequilibrium mapping via cladistic analysis of single-nucleotique polymorphism haplotypes.** *Am J Hum Genet* 2004, **75:**35-43.
30. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cézard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G: **Association of NOD2 leucin-rich repeat variants with susceptibility to Crohn's disease.** *Nature* 2001, **411:**599-603.
31. Lazzeroni LC: **Linkage disequilibrium and gene mapping: an empirical least-square approach.** *Am J Hum Genet* 1998, **62:**159-170.
32. McPeek MS, Strahs A: **Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale genetic mapping.** *Am J Hum Genet* 1999, **65:**858-875.
33. Rannala B, Reeve JP: **High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence.** *Am J Hum Genet* 2001, **69:**159-178.
34. Zhao H, Pfeiffer R, Gail MH: **Haplotype analysis in population genetics and association studies.** *Pharmacogenomics* 2003, **4(2):**171-178.
35. Niu T: **Algorithms for inferring haplotypes.** *Genet Epidemiol* 2004, **27:**.
36. Grassly NC, Rambaut A: **Treevolve, version 1.32.** 2000.
37. Bourgain C, Génin E, Quesneville H, Clerget-Darpoux F: **Search for multifactorial disease susceptibility genes in founder populations.** *Ann Hum Genet* 2000, **64:**255-265.
38. APOE T, Alzheimer Disease meta analysis consortium: **Effects of age, gender and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease.** *J Am Med Assoc* 1997, **278:**1349-1356.
39. Felsenstein J: *Inferring phylogenies, Sinauer Associates, Inc. 2004 chap. Chapter one: Parsimony methods* :1-18. [ISBN:0878931775].
40. Swofford DL: **PAUP Phylogenetic Analysis Using Parcimony. Version 4.0b10.** Sunderland, Massachusetts: Sinauer Associates; 2002.
41. Daly MJ, Kruglyak L, Pratt S, Houstis N, Reeve MP, Kirby A, Lander ES: **Genehunter, version 2.1.** 2001.
42. Vermeire S, Wild G, Kocher K, Cousineau J, Dufresne L, Bitton A, Langelier D, Pare P, Lapointe G, Cohen A, Daly MJ, Rioux J: **CARD15 genetic variation in a Quebec population: prevalence, genotype-phenotype relationship, and haplotype structure.** *Am J Hum Genet* 2002, **71:**74-83.
43. Becker T, Knapp M: **A powerful strategy to account for multiple testing in the context of haplotype analysis.** *Am J Hum Genet* 2004, **75(4):**561-570.
44. Abecasis GR, Cookson WO: **GOLD-graphical overview of linkage disequilibrium.** *Bioinformatics* 2000, **16(2):**182-183.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

3.4 Les facteurs influençant la méthode

Nous avons étudié l'influence de différents facteurs sur l'efficacité de la méthode `ALTree` . Pour comparer les puissances de détection de l'association dans des conditions différentes, les p-values doivent être identiques. Le processus de simulation que nous avons utilisé dans l'article 2 n'est donc pas satisfaisant et donne des résultats parfois difficilement interprétables. Théoriquement, la comparaison pourrait être effectuée en utilisant les p-values évaluées par permutation. Mais, en pratique, le temps de calcul est tel (environ 48h de calcul sur un Pentium III, 930 MHz, 512 Mo de RAM) qu'il n'est pas envisageable d'utiliser ce procédé en simulations. Pour ces raisons, dans cette partie, nous avons principalement étudié l'efficacité de la localisation des sites de susceptibilité. Nous ne reviendrons sur la puissance de la détection de l'association que brièvement dans le dernier paragraphe (page 102).

3.4.1 La taille de l'échantillon

Nous avons étudié l'influence de la taille de l'échantillon en réalisant des simulations à partir des données de la base SeattleSNPs en utilisant le processus de simulation décrit page 71. Nous avons travaillé sur les données du gène *PLAU*. Les résultats de ces simulations sont récapitulées dans la Tableau 3.3 et sur la Figure 3.4.

taille ech.[a]	bonne détection[b]	aucune détection[b]	mauvaise détection[b]
50 + 50	33,0	22,0	45,0
100 + 100	57,6	12,8	29,6
200 + 200	78,9	3,6	16,6
300 + 300	91,1	0,6	8,3
400 + 400	95,9	0,9	3,2
500 + 500	97,5	0,1	2,4
1000 + 1000	99,9	0,0	0,1

[a] taille de l'échantillon : nombre de malades + nombre de témoins
[b] les valeurs correspondent au pourcentage de réplicats pour lesquels :
- le bon site a été identifié, soit de façon unique, soit parmi un ensemble de sites ("bonne détection") ;
- aucun site n'a été identifié, c'est-à-dire que tous les V_i sont négatifs ou nuls ("aucune détection") ;
- un ou plusieurs sites différents du site de susceptibilité simulé ont été identifiés ("mauvaise détection")

Les valeurs ont été obtenues par l'analyse de 1000 réplicats sur le gène *PLAU*.

TAB. 3.3: Influence de la taille de l'échantillon sur l'efficacité de localisation

Le SNP de susceptibilité est le SNP 1 ($f = 0,304$). Les pénétrances valent $0,01$ pour les homozygotes non porteurs de l'allèle à risque, $0,02$ pour les hétérozygotes et $0,1$ pour les homozygotes porteurs. Les données permettent de reconstruire par parcimonie un unique arbre phylogénétique non enraciné. Les états de caractères sont optimisés avec l'option `acctran`.

Les résultats sont tout à fait conformes à nos attentes : lorsque la taille de l'échantillon augmente, l'efficacité de la localisation du site de susceptibilité augmente. Le Tableau 3.3 montre aussi que, dès que la taille de l'échantillon dépasse 50 malades et 50 témoins, le

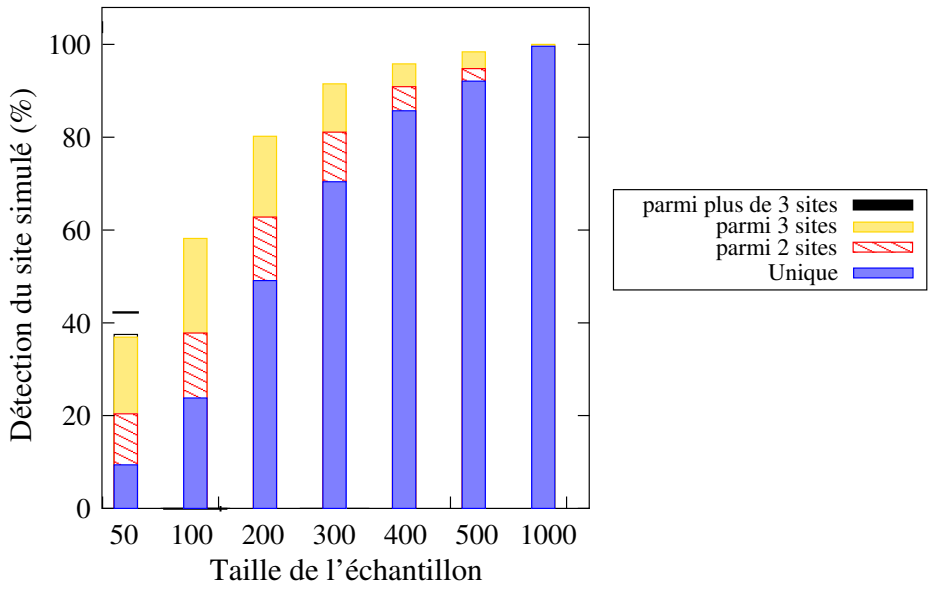


FIG. 3.4: Influence de la taille de l'échantillon sur le nombre de sites détectés avec le site de susceptibilité

Cette figure représente, pour chaque taille d'échantillon, le pourcentage de réplicats, sur 1000, parmi lesquels le site de susceptibilité est identifié de façon unique (1), ou avec un autre site (2), ou avec deux autres sites (3), etc.
Les paramètres des simulations sont les mêmes que celles du Tableau 3.3.

nombre de réplicats dans lesquels le site de susceptibilité simulé est identifié est supérieur au nombre de réplicats dans lesquels un ou plusieurs "mauvais" sites sont détectés. Il montre aussi qu'il y a plus de réplicats lorsqu'un "mauvais" site est identifié que de réplicats lorsqu'aucun site n'est détecté. La Figure 3.4 montre de plus que, globalement, le site de susceptibilité est détecté parmi un nombre réduit de sites possédant le même V_i. Enfin, on note que plus la taille de l'échantillon est élevée, plus le site est identifié avec précision car il est alors plus souvent le seul site détecté.

3.4.2 Le modèle maladie

Dans l'article 2, nous avons montré que l'efficacité d'`ALTree` est différente pour les deux modèles maladies dans lesquels deux sites de susceptibilité sont impliqués. Dans cette section, nous allons étudier plus précisément l'influence de deux paramètres du modèle maladie : la fréquence de l'allèle à risque et la pénétrance. Les simulations sont réalisées comme dans le paragraphe précédent en utilisant les données de la base Seattle SNPs. Nous avons travaillé sur deux gènes différents : *PLAU* et *IL13*.

3.4.2.1 La fréquence de l'allèle de susceptibilité

Pour étudier l'influence de la fréquence de l'allèle de susceptibilité sur l'efficacité de la localisation, nous avons tour à tour considéré chaque allèle à chacun des marqueurs du gène

comme étant l'allèle de susceptibilité. Pour le gène *IL13* qui comporte 12 marqueurs, nous avons donc 24 fréquences possibles pour l'allèle de susceptibilité. En réalité, comme certains sites sont en déséquilibre de liaison complet, nous n'observons que 16 fréquences différentes. De même, pour le gène *PLAU* qui comporte 16 marqueurs, nous disposons de 32 fréquences dont seules 22 sont différentes entre elles. L'efficacité de la localisation est représentée sur la Figure 3.5.

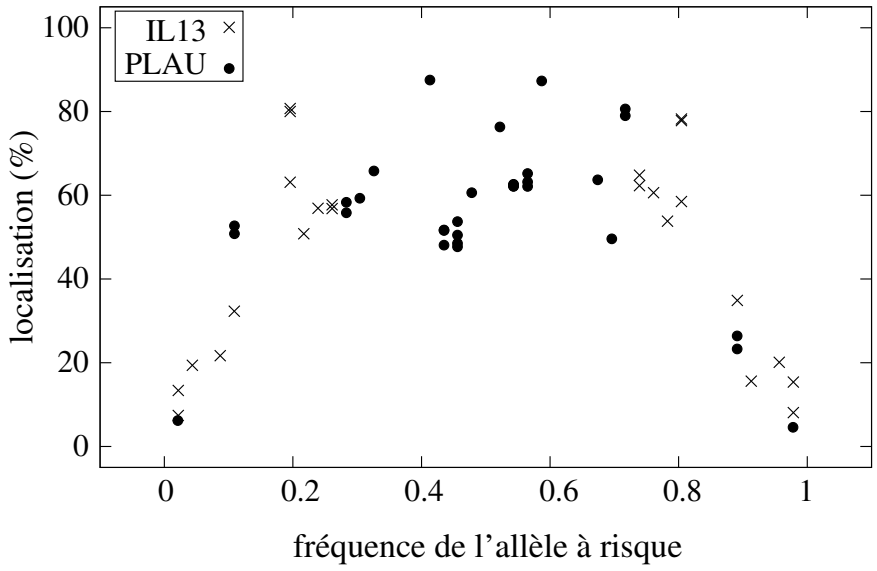


FIG. 3.5: Influence de la fréquence de l'allèle de susceptibilité sur l'efficacité de la localisation

Les résultats sont obtenus sur 1000 réplicats sur les gènes *PLAU* et *IL13*. La taille de l'échantillon est fixée à 100 malades et 100 témoins. Les pénétrances sont respectivement de 0.03 pour les homozygotes non porteurs de l'allèle à risque, 0,06 pour les hétérozygotes et 0,3 pour les homozygotes porteurs de l'allèle à risque. Les données permettent de reconstruire par parcimonie respectivement un unique arbre phylogénétique (gène *PLAU*) et 89 arbres équiparcimonieux (gène *IL13*) non enracinés. Les états de caractères sont optimisés avec l'option `acctran`.

Sur la Figure 3.5, on distingue trois parties :

- lorsque la fréquence de l'allèle à risque est inférieure à 0,2, l'efficacité de localisation augmente avec la fréquence de l'allèle à risque. Ce résultat est conforme à ce que l'on attend du fait de la présence possible de phénocopies[7] : plus la fréquence de l'allèle de susceptibilité est faible, plus le nombre d'individus malades non porteurs de cet allèle est élevé (voir Tableau 3.4), ce qui perturbe la localisation.
- lorsque la fréquence de l'allèle à risque est comprise entre 0,2 et 0,8, l'efficacité de localisation est comprise entre 50% et 80%. Dans cette partie de la courbe, on observe une forte variabilité de l'efficacité pour les différents allèles de susceptibilité, mais elle ne semble pas être corrélée à leur fréquence ;
- lorsque la fréquence de l'allèle à risque est supérieure à 0,8, l'efficacité diminue avec la fréquence de l'allèle de susceptibilité. C'est le phénomène inverse de celui observé

[7] Le taux de phénocopie correspond à la proportion d'individus malades ne portant pas l'allèle à risque. Chez ces individus, la maladie est due à d'autres facteurs, soit génétiques, soit environnementaux

pour les fréquences faibles. Plus l'allèle de susceptibilité est fréquent, plus le nombre d'individus témoins porteurs de cet allèle est élevé, ce qui perturbe la localisation (voir Tableau 3.4).

	NP[a]	HetZ[b]	P[c]	NP[a]	HetZ[b]	P[c]	NP[a]	HetZ[b]	P[c]
	f = 0,02			f = 0,26			f = 0,98		
malades	84	15	1	23	43	28	0	0	100
témoins	97	3	0	57	39	4	0	6	94

Homozygote non porteur de l'allèle à risque
Hétérozygote
Homozygote porteur de l'allèle à risque

TAB. 3.4: Exemple de répartition des individus malades et témoins simulés dans les différentes catégories génotypiques pour différentes fréquences de l'allèle à risque

Ce tableau présente la répartition des 100 individus malades et des 100 individus témoins dans les trois catégories génotypiques. Les données ont été obtenues en utilisant le gène *IL13*. La pénétrance est fixée à 0.03 pour les homozygotes non porteurs, 0.06 pour les hétérozygotes et 0,3 pour les homozygotes porteurs.

Dans le cas des maladies complexes qui sont dues à des polymorphismes fréquents (0,2<f<0,8), la fréquence de l'allèle à risque aura donc peu d'effet sur l'efficacité de la méthode `ALTree` .

3.4.2.2 La pénétrance de la maladie

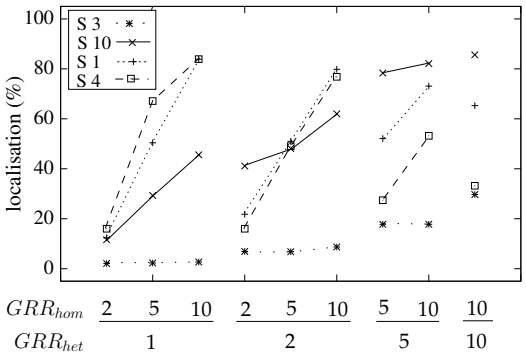
Pour étudier l'influence de la pénétrance sur l'efficacité de la localisation des sites de susceptibilité, nous avons réalisé des simulations sous 9 modèles différents de pénétrance pour différentes valeurs de la fréquence f de l'allèle de susceptibilité. Ces simulations ont été réalisées pour les deux gènes, *PLAU* et *IL13*. La valeur de la pénétrance pour les homozygotes non porteurs de l'allèle à risque a toujours été fixée à 0,01. Aussi, pour la suite, nous utiliserons la notion de risque relatif génotypique[8](GRR pour *genotype relative risk*).

Les résultats des simulations sont représentés sur la Figure 3.6.

La Figure 3.6 montre que l'effet de la pénétrance varie en fonction de la fréquence de l'allèle de susceptibilité.

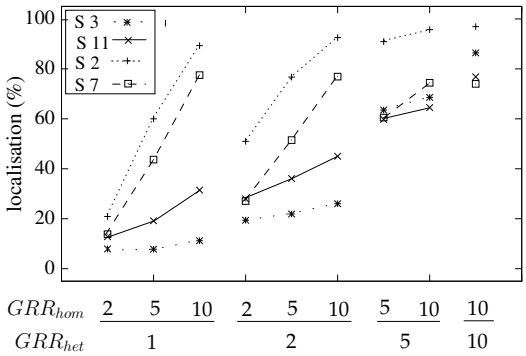
Pour des fréquences faibles (≤ 0.04), l'augmentation des GRR_{hom} pour un GRR_{het} constant influe peu sur l'efficacité de la méthode. Ceci s'explique par le nombre d'homozygotes porteurs de l'allèle à risque très faible, voire nul (voir Tableau 3.5). Lorsque la fréquence de l'allèle de susceptibilité est faible, on observe aussi que l'augmentation des GRR_{het} entraîne une augmentation de puissance relativement importante. La raison en est que les individus malades sont, dans ce cas, essentiellement hétérozygotes ou homozygotes non porteurs de l'allèle à risque. Ainsi, lorsque le risque relatif des hétérozygotes augmente, le nombre de malades hétérozygotes augmente par rapport aux non porteurs, ce qui augmente l'efficacité de la méthode.

[8]Notion définie page 8



site S3 : f=0,02
site S10 : f=0,11
site S1 : f=0,30
site S4 : f=0,54

(a) Gène *PLAU*



site S3 : f=0,04
site S11 : f=0,11
site S2 : f=0,19
site S7 : f=0,26

(b) Gène *IL13*

FIG. 3.6: Influence de la pénétrance de la maladie sur l'efficacité de la localisation

Les résultats sont obtenus sur 1000 réplicats sur les gènes *PLAU* et *IL13*. Les simulations ont été réalisées pour différentes valeurs de la fréquence de l'allèle de susceptibilité. Les numéros des sites de susceptibilité correspondants et les valeurs des fréquences sont précisées sous chaque figure. La taille de l'échantillon est fixée à 100 malades et 100 témoins. Les données permettent de reconstruire par parcimonie respectivement un unique arbre phylogénétique (gène *PLAU*) et 89 arbres équiparcimonieux (gène *IL13*) non enracinés. Les états de caractères sont optimisés avec l'option `acctran`.

		$f = 0,02$[d]			$f = 0,28$[d]		
GRR_{het}[e]	GRR_{hom}[e]	HomZ NP[a]	HetZ[b]	HomZ P[c]	HomZ NP[a]	HetZ[b]	HomZ P[c]
1	2	98	2	0	52	36	12
1	5	95	5	0	41	32	27
1	10	94	6	0	34	23	43
5	5	80	18	2	17	65	18
5	10	86	13	1	9	67	24
10	10	69	29	2	9	74	17

[a] Homozygote non porteur de l'allèle à risque
[b] Hétérozygote
[c] Homozygote porteur de l'allèle à risque
[d] Fréquence de l'allèle de susceptibilité
[e] Risque relatif génotypique des homozygotes porteurs et des hétérozygotes

TAB. 3.5: Exemple de répartition des individus malades simulés dans les différentes catégories génotypiques pour différents modèles de maladie

Ce tableau représente le nombre d'individus obtenu dans chacune des catégories génotypiques lors de la simulation de 100 individus de phénotype malade pour différentes valeurs des pénétrances et de la fréquence de l'allèle à risque. Il s'agit d'un exemple sur une seule simulation. Le gène considéré lors des simulations est le gène IL13.

Pour des fréquences élevées (≥ 0.11), le nombre d'individus homozygotes porteurs de l'allèle à risque est important (voir Tableau 3.5). L'augmentation du GRR_{hom} entraîne alors une augmentation de l'efficacité de localisation du site de susceptibilité. On observe aussi que l'augmentation des GRR_{het} pour des GRR_{hom} constants n'entraîne qu'une augmentation d'efficacité limitée, voire même, pour des fréquences de l'allèle de susceptibilité et des GRR_{hom} très élevées, que cela diminue l'efficacité de la méthode. Quand le GRR_{hom} et la fréquence de l'allèle sont élevés, les individus malades simulés sont majoritairement hétérozygotes et homozygotes porteurs. Lorsque le GRR_{het} augmente et que le GRR_{hom} reste stable, la proportion d'hétérozygotes va augmenter, au détriment des homozygotes porteurs (voir les deux dernières lignes du Tableau 3.5). Or, comme on affecte aux haplotypes le statut maladie de l'individu qui les porte, pour un hétérozygote, il y a nécessairement un des haplotypes pour lequel le statut maladie est faux. L'augmentation du GRR_{het} conduit donc, dans ce cas, à diminuer la puissance de la méthode.

En conclusion, on peut dire que dans le cas où l'allèle à risque est peu fréquent, l'efficacité sera maximale si les deux GRR sont élevés. En particulier, l'efficacité de localisation sera plutôt meilleure pour des modèles à effet dominant. Pour des allèles plus fréquents, les conditions optimales de détection du site de susceptibilité sont plus variables. Lorsque le GRR_{hom} est élevé, l'efficacité sera meilleure pour des modèles à effets récessifs ou intermédiaires.

3.4.3 Les paramètres liés à la reconstruction phylogénétique

La méthode `ALTree` repose sur la reconstruction d'une phylogénie. Les conditions de cette reconstruction peuvent donc avoir une influence sur l'efficacité de la méthode. Nous nous sommes intéressés à trois paramètres : la méthode de reconstruction de l'arbre, l'enracinement et les méthodes d'optimisation des caractères.

3.4.3.1 La méthode de reconstruction de l'arbre

En utilisant le logiciel `TREEVOLVE` , nous avons simulé 5 ensembles de 24 haplotypes différents sous un modèle de coalescence. Ces ensembles d'haplotypes correspondent à 5 topologies d'arbre-guide différentes ($T1$ à $T5$) simulées par `TREEVOLVE` . Chacun de ces ensembles d'haplotypes sert de base pour simuler et analyser 1000 réplicats selon la méthode décrite page 71. Pour chaque topologie, nous avons comparé l'efficacité d'`ALTree` lorsque la phylogénie est reconstruite par une méthode de parcimonie et par une méthode de maximum de vraisemblance. Les conditions de la reconstruction de l'arbre sont celles décrites page 73.

Les résultats des simulations sont représentés dans le Tableau 3.6

Les résultats des simulations montrent que l'efficacité est particulièrement élevée pour quatre des cinq jeux de données, lorsque la méthode de parcimonie est utilisée. Cette observation est probablement due à une surestimation de l'efficacité de la méthode de parcimonie par rapport à celle de maximum de vraisemblance. En effet, dans nos simulations, nous avons analysé plusieurs arbres équiparcimonieux alors que seul l'arbre le plus vraisemblable est pris en compte pour l'analyse par maximum de vraisemblance. Cette dernière stratégie n'est justifiée que si la vraisemblance maximale est nettement supérieure aux vraisemblances de tous les autres arbres ou si tous les arbres dont la vraisemblance est proche conduisent à

Topologie de l'arbre guide	$T1$	$T2$	$T3$	$T4$	$T5$
Nb arbres équiparc.	5	30	1	128[a]	142[a]
Parcimonie	**92.1**	**90.6**	81.5	**87.2**	**72.0**
ML	67.6	63.3	**90.2**	84.5	63.0

[a] Pour des raisons de temps de calcul, seuls 50 arbres choisis aléatoirement sont étudiés.
L'efficacité est exprimée en pourcentage de réplicats dans lesquels le site de susceptibilité a été identifié (seul ou parmi un ensemble d'autres sites).

TAB. 3.6: Influence de la méthode de reconstruction de l'arbre sur l'efficacité de localisation du site de susceptibilité

Les résultats sont obtenus sur 1000 réplicats. Les pénétrances valent 0,03 pour les homozygotes non porteurs de l'allèle à risque, 0,06 pour les hétérozygotes et 0,3 pour les homozygotes porteurs. La fréquence de l'allèle à risque vaut 0.2. La taille de l'échantillon est fixée à 100 malades et 100 témoins. Les arbres ne sont pas enracinés.

des efficacités similaires lors de l'analyse avec `ALTree` . Nous avons alors vérifié si c'était le cas dans nos données en prenant l'exemple de la topologie $T5$.

Nous avons tout d'abord étudié la distribution des vraisemblances pour les arbres les plus vraisemblables. Comme le logiciel `PHYML` permet uniquement d'obtenir l'arbre le plus vraisemblable, pour obtenir les vraisemblances de plusieurs arbres, nous avons calculé les vraisemblances des 142 arbres équiparcimonieux obtenus avec `PAUP` car nous supposons que ces arbres font partie des arbres les plus vraisemblables. La distribution des vraisemblances des 142 arbres est décrite dans le Tableau 3.7.

min[a]	max[a]	médiane	moyenne	écart-type
140,044	143,313	140,527	140,877	0,765

[a] Vraisemblances minimale et maximale parmi les 142
Les valeurs correspondent à l'opposé du logarithme de la vraisemblance.

TAB. 3.7: Caractéristiques des vraisemblances des 142 arbres les plus parcimonieux (topologie $T5$)

Ces résultats montrent que les vraisemblances sont très proches et groupées près de la vraisemblance maximum. Nous avons alors recherché si l'efficacité d'`ALTree` observée pour les 3 arbres les plus vraisemblables était équivalente. Nous avons observé que pour ces trois arbres, l'efficacité de localisation du site de susceptibilité d'`ALTree` est respectivement de 25,9%, 74,6% et 72,0%. Elles sont donc très différentes.

Finalement, comme il existe plusieurs arbres dont la vraisemblance est égale ou très proche du maximum et que l'efficacité de la localisation du site de susceptibilité avec `ALTree` est très variable sur ces arbres, il n'est donc pas du tout justifié de ne travailler que sur l'arbre de vraisemblance maximale alors qu'on considère tous les arbres équiparcimonieux. En effet, lorsqu'on ne considère qu'un seul arbre, il est possible qu'il ne permette pas de localiser efficacement le site de susceptibilité alors que d'autres arbres, équivalents

en terme de parcimonie ou dont les vraisemblances sont très proches, le permettent. Le Tableau 3.8 illustre d'ailleurs ce propos dans le cas de l'utilisation de la parcimonie : plus le nombre d'arbres équiparcimonieux pris en compte dans l'analyse avec `ALTree` est grand, meilleure est l'efficacité de la localisation.

Nb arbres équiparc. (jeu de données n° 5)	1	10	50	142[a]
Efficacité (%)	61.1	69.8	72,0	84,0

[a] Tous les arbres équiparcimonieux sont analysés

TAB. 3.8: Influence du nombre d'arbres équiparcimonieux utilisés par `ALTree`

Afin de réellement comparer les performances des méthodes de parcimonie et de maximum de vraisemblance, il faut donc considérer le même nombre d'arbres pour les deux méthodes. Seule la topologie $T3$ le permet car il n'y a alors qu'un seul arbre le plus parcimonieux. On observe dans ce cas que l'efficacité d'`ALTree` est meilleure par la méthode de maximum de vraisemblance (90,2% contre 81,5% avec la méthode de parcimonie).

En conclusion, pour certaines topologies d'arbres et lorsque l'on n'analyse qu'un seul arbre, l'utilisation du maximum de vraisemblance peut se révéler plus efficace que la parcimonie. Ces observations laissent présager que l'utilisation d'un ensemble d'arbres dont la vraisemblance est proche du maximum pourrait donner de bons résultats.

3.4.3.2 Comparaison des méthodes de parcimonie et de vraisemblance pour l'optimisation des caractères

En parcimonie, il y a deux possibilités d'optimiser les caractères : `acctran` et `deltran`[9] alors qu'en maximum de vraisemblance, les probabilités des états aux nœuds sont définies sans ambiguïté, et nous avons choisi d'affecter à chaque nœud, les états de caractères les plus probables. Nous nous sommes alors interrogés sur l'efficacité respective de ces trois méthodes d'optimisation des caractères.

Pour répondre à cette question, nous avons analysé 11 arbres parmi les 142 équiparcimonieux, trois d'entre eux sont les arbres les plus vraisemblables, les autres sont des arbres choisis aléatoirement. Les résultats de ces simulations sont présentés dans le Tableau 3.9.

Ces résultats confirment qu'il existe une très forte variabilité de l'efficacité de localisation du site de susceptibilité en fonction de l'arbre qui est analysé :

- entre 28,5% et 78,3% de localisation pour l'analyse par parcimonie en utilisant `deltran`
- entre 23,2% et 54,1% de localisation pour l'analyse par parcimonie en utilisant `acctran;`
- entre 19,1% et 74,6% de localisation pour l'analyse par maximum de vraisemblance.

Ils montrent aussi qu'il existe un effet de la méthode d'optimisation des caractères mais que cet effet varie en fonction de l'arbre étudié. Globalement, sur ces données, `acctran` semble donner de moins bons résultats que les deux autres méthodes. Pour les 3 arbres de vraisemblance maximum, `deltran` et la méthode maximum de vraisemblance sont à peu près équivalentes, avec un léger avantage pour la méthode de maximum de vraisemblance. En revanche, pour les autres arbres, l'efficacité peut être très différente. On peut aussi noter

[9] `acctran` et `deltran` ont été définis page 26

		Parcimonie				Maximum vraisemblance	
		`deltran`		`acctran`			
n° de l'arbre	-log(L)	détection[a]	unique[b]	détection[a]	unique[b]	détection[a]	unique[b]
ARBRES LES PLUS VRAISEMBLABLES							
9	140.04	**29,1**	29,1	28,2	27,8	25,9	25,4
31	140.04	68,3	65,1	28,4	24,0	**74,6**	72,0
67	140.04	67,7	64,7	26,3	25,9	**68,3**	65,1
AUTRES ARBRES							
1	140.32	30,1	30,1	39,8	38,9	**52,5**	49,4
19	140.35	28,5	28,5	**34,8**	33,8	19,1	18,5
89	140.42	69,0	65,4	32,3	31,5	**74,2**	70,7
42	140.53	**68,8**	67,5	25,6	25,6	29,3	28,2
119	141.19	**78,3**	73,5	54,1	51,2	24,6	23,3
102	141.27	**73,5**	71,9	23,3	22,9	29,6	28,4
142	141.65	**74,3**	71,5	23,2	22,3	25,7	25,0
140	143.21	68,5	65,1	49,9	48,9	**71,8**	68,7

[a] Efficacité d'`ALTree` pour localiser le site de susceptibilité seul ou parmi un ensemble d'autres sites

[b] Efficacité d'`ALTree` pour localiser le site de susceptibilité sans ambiguïté

TAB. 3.9: Efficacité d'`ALTree` pour différentes méthodes d'optimisation des caractères

Les simulations ont été réalisées sur 11 arbres parmi les 142 arbres équiparcimonieux obtenus pour la topologie $T5$. La taille de l'échantillon simulé est 100 malades et 100 témoins. Les pénétrances sont 0,03 pour les homozygotes non porteurs, 0,06 pour les hétérozygotes et 0,3 pour les homozygotes porteurs. La fréquence de l'allèle de susceptibilité est de 0,21.

qu'en fonction de l'arbre étudié, les résultats obtenus en utilisant la méthode de maximum de vraisemblance sont plus proches de ceux obtenus avec `acctran` (arbres 42, 102, 142) ou avec `deltran` (arbres 31, 67, 89, 140). Il n'est donc pas possible de tirer une conclusion générale sur l'efficacité relative des trois méthodes à partir de l'observation de ces résultats.

3.4.3.3 L'enracinement et les méthodes d'optimisation des caractères

Les résultats de toutes les simulations précédentes ont été obtenus en travaillant sur un arbre non enraciné. Il est cependant possible de travailler sur des arbres enracinés en faisant des hypothèses concernant les états ancestraux des caractères. Nous avons donc étudié si l'apport d'information concernant les états ancestraux permet d'améliorer l'efficacité d'`ALTree` en nous concentrant sur le cas de l'analyse par parcimonie.

De plus, au paragraphe précédent, nous avons observé une très forte variabilité entre les résultats obtenus avec différentes méthodes d'optimisation des caractères. Dans ce paragraphe, nous nous sommes à nouveau intéressés à ce problème, mais cette fois-ci, en lien avec l'enracinement de l'arbre et la fréquence de l'allèle de susceptibilité. Pour cela, nous avons évalué l'efficacité de localisation du site de susceptibilité[10] en réalisant des simulations

[10] Les simulations de l'article 2 ont montré que le mode d'optimisation des caractères n'avait pas d'influence sur la puissance de détection de l'association. Ce résultat était attendu car la méthode de détection de l'associa-

dans quatre conditions (arbre enraciné ou non et optimisation des caractères avec `acctran` ou `deltran`) pour des fréquences différentes de l'allèle de susceptibilité. Nous avons utilisé les données des gènes *IL13* et *PLAU*.

Intérêt de l'enracinement Les résultats du Tableau 3.10 montrent que l'importance de l'enracinement dépend en partie des fréquences de l'allèle de susceptibilité. En effet, on observe que pour des fréquences très faibles (<0,1), l'enracinement permet toujours d'augmenter l'efficacité de la localisation, quelle que soit la méthode d'optimisation des caractères utilisée. Au contraire, lorsque les fréquences sont fortes (>0,76), l'enracinement tend à diminuer l'efficacité de la localisation. Lorsque les fréquences sont intermédiaires, selon les locus de susceptibilité considérés et la méthode d'optimisation des caractères, l'enracinement permet ou non d'augmenter l'efficacité de la méthode.

Influence de la méthode d'optimisation des caractères Les résultats du Tableau 3.10 montrent que la méthode d'optimisation a une très grande influence sur l'efficacité de la localisation. Pour certaines fréquences du locus de susceptibilité, l'efficacité de la localisation varie d'un facteur deux entre les deux méthodes d'optimisation des caractères.

Lorsque l'arbre n'est pas enraciné, il n'y a pas de relation claire entre la méthode d'optimisation des caractères, la fréquence de l'allèle de susceptibilité et l'efficacité de la localisation.

Lorsque l'arbre est enraciné, l'efficacité de localisation est globalement meilleure si les caractères sont optimisés avec `deltran` pour des fréquences de l'allèle à risque inférieures à 0,5. Pour des fréquences de l'allèle à risque supérieures à 0,5, on observe en revanche que l'efficacité de la localisation est globalement meilleure avec `acctran`. Si l'on ne considère que les fréquences extrêmes (inférieures à 0,1 et supérieures à 0,9), cette observation est encore plus nette. La moindre efficacité d'`acctran` pour des fréquences de l'allèle de susceptibilité inférieures à 0,5 a déjà été notée dans l'article 2. Dans ces simulations, nous avons montré que `deltran` donne de meilleurs résultats pour 8 jeux de données sur les 10 analysés, que les deux méthodes sont équivalentes pour un jeu de données et qu'`acctran` est plus efficace pour un jeu de données. Lorsque la fréquence de l'allèle à risque est relativement faible, peu d'haplotypes vont le porter. On s'attend donc à ce que l'allèle de susceptibilité apparaisse plutôt tardivement dans l'arbre (vers les feuilles). En cas d'ambiguïté d'état du caractère S chez certains haplotypes (états " ?"), l'option `deltran` qui place les changements d'état le plus tardivement possible dans l'arbre aura alors tendance à permettre une meilleure détection de site de susceptibilité. Ce phénomène est illustré par la Figure 3.7.

Le même genre de phénomène permet d'expliquer les meilleurs résultats d'`acctran` pour des fréquences de l'allèle de susceptibilité supérieures à 0,75. Il faut cependant noter que la fréquence de l'allèle de susceptibilité n'est pas le seul facteur qui influe sur l'efficacité relative des deux méthodes d'optimisation des caractères. La position des changements d'état du site de susceptibilité dans l'arbre peut jouer un grand rôle : certaines configurations seront plus favorablement détectées par l'une ou l'autre des méthodes d'optimisation. Cela explique les variations parfois très fortes entre les résultats obtenus pour des fréquences voisines.

---

tion ne fait pas intervenir la reconstruction des caractères sur les arbres.

		sans racine		avec racine	
gène	fréquence	deltran	acctran	deltran	acctran
IL13	0,02	**23,8**	7,4	**42,8**	9,0
IL13	0,02	**24,6**	13,4	**50,1**	21,2
IL13	0,04	**24,5**	19,4	**39,0**	35,3
IL13	0,09	**36,1**	21,7	**43,6**	18,2
IL13	0,11	22,5	**32,3**	**31,1**	20,5
PLAU	0,11	**51,1**	37,9	52,6	**52,7**
IL13	0,19	54,6	**80,0**	**77,1**	69,4
IL13	0,19	25,9	**63,1**	**67,4**	25,4
IL13	0,22	30,4	**50,8**	**51,0**	27,7
IL13	0,24	31,8	**56,9**	**61,0**	33,3
IL13	0,26	**76,3**	57,7	60,7	**89,4**
PLAU	0,28	**58,8**	37,9	25,1	**55,7**
PLAU	0,33	66,4	**72,8**	**68,4**	66,0
PLAU	0,41	**84,4**	51,0	50,6	**87,2**
PLAU	0,43	51,1	**55,2**	**52,6**	50,2
PLAU	0,46	49,2	**71,2**	**69,3**	49,0
PLAU	0,48	59,8	**93,5**	**77,6**	57,7
PLAU	0,52	61,6	**70,3**	70,4	**75,6**
PLAU	0,54	46,1	**57,2**	52,5	**64,8**
PLAU	0,56	**51,0**	42,1	40,6	**63,9**
PLAU	0,59	**77,4**	40,1	36,9	**86,1**
PLAU	0,67	46,6	**57,8**	46,0	**61,8**
PLAU	0,72	**58,4**	27,7	27,1	**80,8**
IL13	0,74	**64,6**	62,3	25,4	**65,5**
IL13	0,76	24,3	**60,6**	**30,8**	26,5
IL13	0,78	21,3	**53,8**	**41,9**	26,3
IL13	0,80	23,6	**77,8**	**41,7**	21,9
IL13	0,80	41,7	**78,2**	30,4	**47,4**
PLAU	0,89	**58,4**	25,0	23,2	**24,3**
IL13	0,89	11,7	**34,9**	9,9	**24.3**
IL13	0,91	10,2	15,6	4,6	**18,7**
IL13	0,96	**20,3**	**20,1**	0,7	**11,5**
IL13	0,98	**18,8**	8,1	2,1	**7,1**
IL13	0,98	**18,1**	15,4	0,9	**10,8**

Les valeurs sont données en pourcentage de localisation de site de susceptibilité (seul ou parmi plusieurs autres sites)

TAB. 3.10: Influence de l'enracinement et du mode d'optimisation des caractères sur l'efficacité de la localisation avec `ALTree`

Les résultats sont obtenus sur 1000 réplicats. Dans chaque réplicat, l'analyse porte sur les 89 arbres équiparcimonieux obtenus pour le gène *IL13* ou sur l'arbre le plus parcimonieux obtenu pour le gène *PLAU*. Les pénétrances valent 0,03 pour les homozygotes non porteurs de l'allèle à risque, 0,06 pour les hétérozygotes et 0,3 pour les homozygotes porteurs. La taille de l'échantillon est fixée à 100 malades et 100 témoins.

La meilleure efficacité pour chaque méthode d'enracinement est écrite en gras.

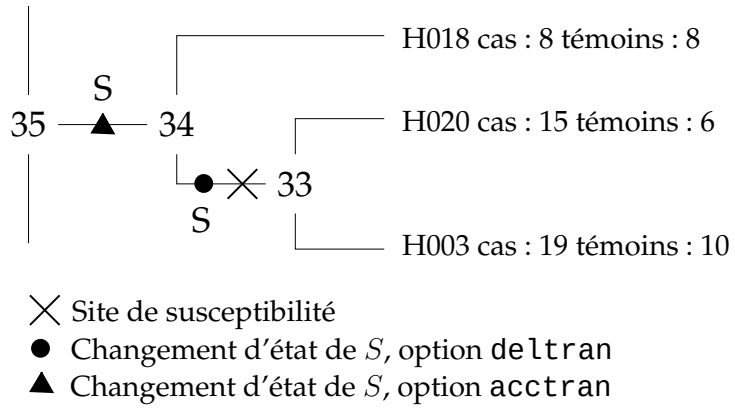


FIG. 3.7: Influence de la méthode d'optimisation des caractères dans le cas où l'allèle de susceptibilité a une fréquence f faible ($f = 0,2$)

D'après le nombre de malades et de témoins dans les différentes branches, le caractère S prendra comme valeur M (majorité de malades) pour H020 et H003, et ? (état indéterminé) pour H018. Avec l'option `acctran`, le changement d'état de S sera le plus haut possible dans l'arbre, c'est-à-dire dans la branche entre les nœuds 35 et 34. Avec l'option `deltran`, le changement d'état de S sera dans la branche entre les nœuds 34 et 33 et le site de susceptibilité pourra être détecté.

3.4.4 Variabilité liée aux jeux de données

Dans l'article 2, nous avons réalisé des simulations pour 10 jeux de données correspondant à 10 topologies d'arbres différentes simulées par `TREEVOLVE` . Les résultats des simulations décrites dans cet article montrent qu'il existe une très forte variabilité entre les résultats obtenus par les différentes méthodes sur les 10 jeux de données alors que tous les autres paramètres de simulation (fréquence de l'allèle de susceptibilité, pénétrance, taille de l'échantillon, etc.) sont identiques. Cette variabilité s'observe à la fois pour la puissance de détection de l'association et pour l'efficacité de la localisation. Nous avons alors recherché quelles pouvaient être les causes de cette variabilité pour la méthode locus par locus (notée SbST) et pour les deux méthodes basées sur des phylogénies : `ALTree` et `CLADHC` .

Le Tableau 3.11 rappelle les valeurs de l'écart-type de la puissance de détection de l'association pour les 10 jeux de données.

Pour la méthode locus par locus : L'écart-type entre les puissances de détection de l'association des 10 jeux de données est très faible lorsque le site de susceptibilité est génotypé. Lorsque le site de susceptibilité n'est pas génotypé, cet écart-type est alors très élevé. Dans ce dernier cas, c'est l'existence d'un site en fort déséquilibre de liaison avec le site de susceptibilité qui va influer sur la puissance de la méthode comme le montrent les résultats de la Figure 3.8a. Cette figure montre l'influence du déséquilibre de liaison maximum entre le site de susceptibilité et les autres marqueurs, à la fois sur la puissance de détection de l'association dans le cas où le site de susceptibilité n'est pas génotypé, et sur l'efficacité de la localisation. On observe que la puissance de détection de l'association augmente lorsque qu'il existe un site en fort déséquilibre. Dans ce cas, en l'absence du site de susceptibilité, c'est ce site en déséquilibre de liaison qui va permettre de détecter l'association. Il n'est donc pas surprenant que l'association soit mieux détectée lorsque le déséquilibre est fort. À l'in-

Modèle	ALTree	CLADHC	SbST[a]
A	5.63	3.24	0.13
B	6.71	5.2	9.81
C	2.24	1.83	1.58
D	11.07	7.84	5.94

[a] Test locus par locus

Les 4 modèles :

A : locus de susceptibilité génotypé.

B : locus de susceptibilité non génotypé

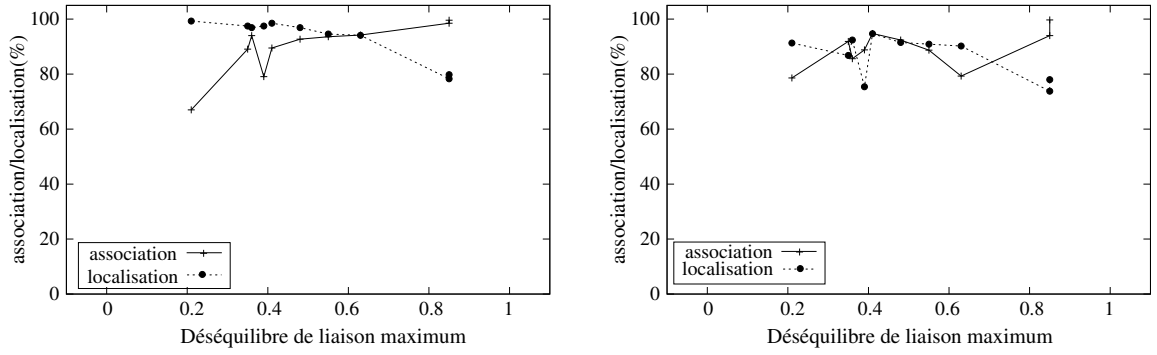
C : deux locus de susceptibilité génotypés, pénétrance des homozoygotes porteurs des deux allèles à risque fort (0,9)

D : deux locus de susceptibilité génotypés, pénétrance des homozoygotes porteurs des deux allèles à risque faible (0,6)

Le détail des modèles est donné page 75

TAB. 3.11: Ecarts-types de la puissance de détection de l'association obtenus lors de l'analyse de 10 jeux de données simulés sous 4 modèles différents

verse, l'efficacité de localisation du site de susceptibilité diminue lorsque le déséquilibre de liaison maximum augmente. En effet, lorsque deux sites sont en très fort déséquilibre de liaison, il est très difficile de distinguer leurs effets respectifs. Ainsi, s'il existe un ou plusieurs sites en fort déséquilibre avec le site de susceptibilité, dans certains réplicats, ces sites seront détectés à la place du site de susceptibilité, ce qui explique la moindre efficacité d'ALTree .



(a) Méthode locus par locus

(b) Méthode ALTree

FIG. 3.8: Influence du déséquilibre de liaison maximum entre le site de susceptibilité et les autres marqueurs

Pour les méthodes phylogénétiques : la différence entre l'écart-type obtenu pour les simulations où le site de susceptibilité est génotypé et pour celle où il ne l'est pas n'est pas aussi importante que pour la méthode locus par locus. Les résultats de la Figure 3.8b confirment que le déséquilibre de liaison n'est pas un facteur principal de variation. Il est

plus probable que les variations observées entre les jeux de données sont dues à des variations de la topologie des arbres reconstruits à partir des haplotypes. En effet, Seltman *et al.* (2001) ont déjà montré que la position dans l'arbre phylogénétique des haplotypes portant l'allèle à risque a une influence sur la puissance de détection de l'association. Elle serait plus élevée lorsque les haplotypes portant l'allèle de susceptibilité sont regroupés en un unique clade. Nos simulations nous ont permis de confirmer ces observations et de les étendre au cas de la localisation des sites de susceptibilité. Nous avons pu déterminer certaines conditions plus ou moins favorables pour la localisation des sites de susceptibilité. En particulier, nous avons fait les trois observations suivantes :

- la présence de petits effectifs dans les branches proches de celles portant les changements d'état des sites de susceptibilité rend la localisation des sites de susceptibilité plus difficile ;
- la présence de réversions peut diminuer l'efficacité de localisation car les haplotypes porteurs de l'allèle de susceptibilité ne forment plus un clade (du fait de la réversion, tous les descendants de l'ancêtre commun porteurs de l'allèle de susceptibilité ne le portent plus). Le jeu de données simulées présentant une réversion est celui qui a la plus faible efficacité de localisation avec `ALTree` (73,8%) ;
- la présence de plusieurs sites qui co-mutent avec le site de susceptibilité ne va pas diminuer l'efficacité de localisation, mais il sera alors impossible d'identifier de façon unique le site réel de susceptibilité.

3.5 Étude de puissance d'une méthode phylogénétique alternative

Une autre méthode phylogénétique a été développée par Durrant *et al.* (2004) et implémentée dans le logiciel `CLADHC` . Les simulations réalisées par Durrant *et al.* (2004) semblent montrer que la méthode est très puissante. Cependant, le processus de simulation utilisé dans leur étude est très particulier (les haplotypes sont fabriqués selon la méthode décrite page 47). De plus, ils ont obtenu des résultats très différents des nôtres. En particulier, dans leur étude, la méthode locus par locus semble vraiment peu puissante (5% pour des risques relatifs de 3 pour les hétérozygotes et 20 pour les homozygotes porteurs de l'allèle de susceptibilité) par rapport aux autres méthodes. Nous avons donc réalisé une nouvelle étude de puissance de la méthode `CLADHC` par des simulations basées sur des haplotypes directement reconstruits à partir de données réelles.

3.5.1 Nouvelle étude de puissance de la méthode de Durrant *et al.*

3.5.1.1 Résumé de l'article 3 : Clustering of haplotypes based on phylogeny : how good a strategy for association testing ?

Par simulations, nous avons évalué la puissance de détection de l'association du test correspondant à la méthode phylogénétique implémentée dans `CLADHC` (Durrant *et al.*, 2004) (nommé T[MAX] par la suite). Nous l'avons aussi comparé à deux autres tests[11] : un test allélique locus par locus et un test haplotypique classique sans regroupement préalable des haplotypes (nommé T[h] par la suite).

[11] Ces tests sont aussi implémentés dans `CLADHC`

Pour cette étude, nous avons réalisé nos simulations à partir de données haplotypiques provenant soit de bases de données disponibles sur internet (SeattleSNPs pour *IL13*, *PLAU* et *TNF*, et HapMap (Consortium, 2003) pour les données du chromosome 20, voir page 71), soit d'une étude de CARD15 dans la maladie de Crohn. Pour chaque jeu de données et pour différentes valeurs de la fréquence de l'allèle à risque et des pénétrances, nous avons formé 1000 réplicats constitués de 1000 malades et de 1000 témoins à partir des haplotypes selon la méthode décrite au page 71. Nous les avons ensuite analysés avec le logiciel `CLADHC` . Contrairement aux simulations de Durrant *et al.* (2004), nous avons conservé les haplotypes en entier, ce qui nous permet d'avoir un déséquilibre de liaison réaliste entre les marqueurs.

Dans nos simulations, nous avons fait varier la fréquence de l'allèle de susceptibilité et le risque relatif des homozygotes porteurs de l'allèle à risque. Nous avons montré que, pour les trois tests, la puissance augmente lorsque la fréquence de l'allèle de susceptibilité augmente et lorsque le risque relatif des homozygotes porteurs de l'allèle à risque augmente. De plus, lorsque le site de susceptibilité n'est pas génotypé, on note que la puissance des trois tests augmente lorsque le déséquilibre de liaison maximum entre le site de susceptibilité et les autres sites (LD_{max}) augmente.

Dans cet article, nous avons aussi souligné la différence entre les résultats que nous avons obtenus et ceux de Durrant *et al.* (2004). Pour des valeurs similaires des risques relatifs et de la fréquence de l'allèle de susceptibilité, nous obtenons des puissances beaucoup plus élevées qu'eux. Cela peut s'expliquer par la longueur des séquences analysées. Durrant *et al.* travaillent sur une séquence de 10 Mb (5216 marqueurs) et, même s'ils ne précisent pas dans leur article le nombre exact de marqueurs utilisés pour l'étude de puissance, on peut supposer que ce nombre est très élevé. Or, les deux études utilisent la correction de Bonferroni pour tenir compte des tests multiples. Comme les tests ne sont pas indépendants, elle n'est pas adaptée à cette situation : la correction réalisée est trop forte, et ce, d'autant plus que le nombre de tests réalisés est élevé. Durrant *et al.* (2004) travaillant sur un très grand nombre de marqueurs, dans leurs simulations, le seuil de signification corrigé par Bonferroni sera très élevé, ce qui explique que tous les tests paraissent peu puissants dans leur étude.

Nous avons aussi montré que, globalement, le test locus par locus est le plus puissant quelle que soit la longueur de la séquence. Ces résultats sont contradictoires avec ceux de Durrant *et al.* (2004) qui montrent que le T[MAX] est plus puissant que les deux autres tests. Une hypothèse permettant d'expliquer cette différence est que le processus de simulation de Durrant *et al.* (2004) génère des déséquilibres de liaison très particuliers favorisant le test T[MAX]. Il faut cependant noter que la puissance de l'analyse locus par locus peut être légèrement surestimée. En effet, pour les deux méthodes haplotypiques, les tests réalisés sont extrêmement dépendants car ils sont réalisés dans des fenêtres chevauchantes. La correction de Bonferroni est donc encore moins adaptée pour les deux analyses haplotypiques que pour l'analyse locus par locus. Cependant, cet argument ne saurait expliquer les différences entre nos résultats et ceux de Durrant *et al.* (2004) car la correction de Bonferroni est employée dans les deux études.

En conclusion, nous pouvons dire que le regroupement sur la base d'un arbre reconstruit par une méthode de distances tel qu'il est implémenté dans `CLADHC` n'apporte généralement pas de gain de puissance par rapport à un test d'association locus par locus.

3.5.2 Article 3

Clustering of haplotypes based on phylogeny : how good a strategy for association testing ? (Bardel *et al.*, 2005b)
Claire Bardel, Pierre Darlu et Emmanuelle Génin. Accepté pour publication par l'*European Journal of Human Genetics*.

Clustering of haplotypes based on phylogeny: how good a strategy for association testing?

Claire Bardel, Pierre Darlu * and Emmanuelle Génin *
INSERM U535, Hôpital Paul Brousse, Villejuif, France
* These authors contributed equally to this work

Running title: clustering of haplotypes

Corresponding author:
Claire Bardel
Génétique Epidémiologique et structure des populations humaines
Hôpital Paul Brousse
Bâtiment Leriche
B.P. 1000
94817 Villejuif Cedex
France
phone number: (33) 1 45 59 53 68
fax number: (33) 1 45 59 53 31
e-mail: bardel@vjf.inserm.fr

Key-words: association test, haplotype phylogeny, cladistic analysis.

Abstract

Haplotypes are now widely used in association studies between markers and disease susceptibility locus. However, when a large number of markers are considered, the number of possible haplotypes increases leading to two problems: an increased number of degrees of freedom that may result in a lack of power and the existence of rare haplotypes that may be difficult to take into account in the statistical analysis. In a recent paper, Durrant *et al*[1] proposed a method, `CLADHC`, to group haplotypes based on distance matrices and showed that this could considerably increase the power of the association test as compared to either single locus analysis or haplotype analysis without prior grouping. Although the authors considered different one-disease-locus susceptibility models in their simulations, they did not study the impact of the linkage disequilibrium (LD) pattern and of the susceptibility allele frequency on their conclusions. Here, we show, using haplotype data from five regions of the genome of different lengths and with different LD patterns, that, when a single disease susceptibility locus is simulated, the prior grouping of haplotypes based on the algorithm of Durrant *et al*[1] does not increase the power of association testing except in very particular situations of LD patterns and allele frequencies.

Introduction

A large number of single nucleotide polymorphisms (SNPs) can now be used to look for an association between a disease and a candidate gene. These markers can either be studied one at a time or jointly in haplotypes. The advantages of haplotypic versus single marker methods have been widely debated in the literature: some studies show that haplotypic tests are more powerful[2, 3, 4] whereas others conclude that single site analysis should be preferred[5, 6, 7]. However, the relative power of these two approaches depends on whether the disease contributing SNPs are among the investigated SNPs or not[8, 9], on the number of disease susceptibility sites[9], on the disease susceptibility model and on the type of interactions between disease contributing sites[10, 11]. The number of SNPs that are considered jointly in haplotypes is also an important parameter since the number of haplotypes increases with the number of investigated SNPs and, consequently, increase the degrees of freedom of tests comparing cases and controls, thus reducing their power. Moreover, as some haplotypes would only be carried by a few individuals, there could be statistical problems owing to small sample sizes making difficult the evaluation of their possible effect on the susceptibility. To face this problem, different strategies have been developed for the grouping of haplotypes. The Templeton et al's method[12] consists in building a cladistic phylogenetic tree of the haplotypes and statistically comparing the number of cases and controls carrying haplotypes from the different nested clades. Recently, Durrant *et al*[1] have proposed a different method in which the grouping of haplotypes is based on a distance metric and the association in the different groups is tested by a regression analysis. The method allows the study of large number of SNPs through a sliding window approach and is implemented in the software `CLADHC`. It differs from the Templeton's one[12] since trees are reconstructed by simple group average linkage (clustering) and rooted, instead of being reconstructed by a parsimony method and unrooted. It also allows the analysis of long haplotypes whereas the Templeton's method focuses on few SNPs. Based on simulations, Durrant *et al*[1] showed that their clustering method may considerably increase the power to detect an association with a genomic region including a single disease susceptibility locus as compared to single-site or classical haplotype analysis. To perform these simulations, Durrant *et al*[1] considered the haplotype data observed in Caucasians in a 10Mb region of the chromosome 20. To determine if their conclusions remain valid with different genes and linkage disequilibrium (LD) patterns, we present here the results of simulations using real haplotype data from five different genomic regions.

Material and methods

Description of the data

Data from the Variation Discovery Ressource Project[13] Various genes are sequenced in 23 unrelated European individuals. SNPs are identified within these genes and the most likely haplotypes are reconstructed using `Phase v2.0`[14, 15]. Three genes are studied here:

- IL13 (Interleukine 13): 6919 base pairs (bp) on the chromosome 5q31: 12 bi-allelic

loci are kept, defining 14 different haplotypes;

- PLAU (Plasminogen Activator Urokinase): 9274 bp on the chromosome 10q24: 16 bi-allelic loci are kept, defining 10 different haplotypes;
- TNF (Tumor Necrosis Factor): 4830 bp on the chromosome 6p21.3: 10 bi-allelic loci are kept, defining 6 different haplotypes.

Data from the chromosome 20 (HapMap project) A 500kb region of chromosome 20 sequenced for 30 CEPH trios is randomly chosen (from position 48362908 to 48862907). The most likely haplotypes are reconstructed using `Phase v2.0`[14, 15]. Two different sets of SNPs are studied:

- CHR20_1 (461 kb): 13 randomly chosen SNPs, defining 37 different haplotypes
- CHR20_2 (442 kb): 12 randomly chosen SNPs, defining 38 different haplotypes

Data on the CARD15 region The data set[16] includes 232 families with two affected children and their parents genotyped for 13 SNPs covering 140 kb in the CARD15 region. Haplotypes were reconstructed using `GENEHUNTER 2.0b`[17]. Haplotypes with missing data were removed and 531 control haplotypes (parental haplotypes non-transmitted to the children) were kept for the analysis. These data include 88 different haplotypes.

The pairwise LD (r^2 value) for these six data sets obtained with `GOLD` software[18] are presented as supplementary data.

The three tests

The data were analyzed using the `CLADHC` software[1], kindly provided by Caroline Durrant. This program considers overlapping sliding windows of SNPs across the haplotypes. In each window, three association tests are performed:

- A single-locus allele-based analysis using Pearson's chi-square test.
- An haplotype-based logistic regression without grouping of haplotypes referred to as T[h].
- An haplotype-based logistic regression analysis with clustering of haplotypes: a tree of the haplotypes is reconstructed using a distance method and a statistics is calculated at each level of the tree. The statistics at the level maximizing the evidence of a disease-marker association is then retained. This test is referred to as T[MAX].

For each window, the program also provides a significant threshold calculated using the Bonferroni correction. The single-locus test is corrected for the number of SNPs, T[h] is corrected for the number of windows and T[MAX] is corrected for the number of windows and for the number of level in the tree.

The simulation process

We start by selecting a site as the disease susceptibility (DS) site in the studied gene. In the following, we will assume that the minor allele at this locus is the one that confers the highest risk of disease (DS allele). To generate the genotype of each individual, pairs of haplotypes are randomly sampled with replacement and the disease status is obtained by applying the penetrance f_2, f_1 and f_0 associated to the different genotypes (2, 1 or 0 DS alleles) at the DS locus. The chosen penetrance values correspond to an heterozygote Genotype Relative Risk (GRR) of 1.5 and to homozygote GRRs of 2, 5 and 10 respectively. We keep sampling with replacement until we obtain 1,000 cases and 1,000 controls. This constitutes a replicate and 1,000 such replicates are simulated for each of the studied penetrance vectors. `CLADHC`[1] was applied on the simulated data either with the DS site being kept or removed. The powers of the three tests (single-locus, T[h] and T[MAX]) to detect an association are compared as in Durrant *et al*[1]. Windows of six markers are considered for T[h] and T[MAX] tests. Type I errors were evaluated by simulations under the null hypothesis of no association, assigning identical values of the penetrances to all three genotypes ($f_2 = f_1 = f_0 = 0.50$) for both cases and controls. However, since the number of performed tests vary from one method to another, different type I errors are obtained. Therefore, to compare the powers of the methods, the data were re-analyzed with different nominal values, until the observed type I errors equal 1% for all the three tests.

For IL13, PLAU, TNF, CHR20_1 and CHR20_2, all the SNPs are considered as the susceptibility site in turn, except when two sites are in complete linkage disequilibrium. In this latter case, only one of the two sites is studied. For the CARD 15 region, as the computation time is really longer due to the larger haplotypic diversity in the data set, only 9 SNPs out of the 13 are analyzed.

Results

Results of the power computations are presented in Table 1 and in Figure 1 and 2. In Table 1, the power to detect an associaion is presented for site 13 of the CARD15 region. As expected, the power of the three tests is higher when the homozygote GRR is high, and when the susceptibility site is included in the analysis. The same results are obtained for all the sites on the four genes.

The power of the three tests for an homozygote GRR of 2 is presented in Figure 1 and 2 for different values of DS allele frequency and for different maximum linkage disequilibrium (LD_{max}, based on the r^2 values) between the susceptibility site and another site. Whatever the test, we can see that the power increases when the DS allele frequency and the LD_{max} increase.

The difference in power for the three tests can be tested by a Friedman two-way analysis of variance. As our data sets are heterogeneous, we test separately the data sets corresponding to long sequences (CHR20_1, CHR20_2 and CARD15) and to short sequences (IL13, PLAU and TNF). The two groups are referred to as LS and SS respectively. When the susceptibility site is present (Figure 1), we find that the power of the three tests is statistically different at the 1% level for LS and SS. As expected in this case, the single locus test is more powerful than the two others since the use of haplotypes increases the number

of tests without adding any information. The difference in power between the two haplotypic tests is not significant at the 5% level for both LS and SS. When the susceptibility site is removed (Figure 2), the difference in power between the three tests is significant at the 1% level for LS and at the 5% level for SS. The pairwise comparisons of the three tests show that T[h] and T[MAX] are not significantly different at the 5% level for both LS and SS and that the single locus test is more powerful than the two other tests. The power of the single locus test is particularly high when LD_{max} is high. However, it should be noted that for the haplotypic tests, the results in the overlapping windows are highly correlated. Thus, they may suffer a greater loss of power due to the Bonferroni correction than the single locus test. With a less conservative correction for multiple testing, the difference of power between the single locus test and the haplotypic tests should be reduced.

For an homozygote GRR of 5, the powers are very high for all the three tests (around 100% for 2/3 of the tested sites), and no significant difference is observed between the three tests.

In this study, as in Durrant *et al*'s[1], we assume that haplotypes can be reconstructed without ambiguity. If this is not the case, haplotype uncertainty should be taken into account in the haplotype tests and their power will be reduced. The power of the single locus test will not be affected and thus the difference in power between the tests will be increased.

Discussion

The results obtained in this study turn out to be very different from those published by Durrant *et al*[1], although the same software is used. For the same GRR values, our power estimates are definitively higher than their. A possible explanation may be the large size of their analyzed region (10 Mb region, 5216 markers): although they do not specify the number of markers included in the "flanking region" used to evaluate the power, we can assume that they are numerous and that a strong correction for multiple testing is used, thus decreasing the power of all the three methods. Our results show that Durrant et al's method does not generally lead to a statistically significant gain in power compared to both single locus and T[h] tests, T[MAX] being the most powerful test only for 6 sites out of the 57 sites tested. One must note that CLADHC is designed to analyse long sequences (several Mb), and thus, in our simulations, we may not be under the optimal conditions, especially for the three short sequences IL13, PLAU and TNF. For the CHR20 data for which distances between markers are close to the ones in the example given by Durrant *et al*[1] (17 of their 23 studied markers are in a 600kb region of the CFTR gene) we indeed find that the performance of T[MAX] compared to T[h] and to the single locus test is better than in the other studied regions. However, even then, T[MAX] has the highest power for only five sites out of 25. For the remaining 20 sites, there is a lack of power when using T[MAX] instead of T[h] or a single locus test. The relative power of the three tests seems to vary with the LD_{max} and with the frequency of the DS allele: all sites but one for which T[MAX] performs better than the two other tests have moderate LD_{max} (<0.6) and moderate frequency of the DS allele (<0.27). The difference between our results and Durrant et al's may be explained by the simulation process they use to generate their sample of haplotypes: they are obtained after a particular processing of

the observed data that might lead to extreme LD patterns. In our simulations, whole haplotypes are sampled from real data sets, thus keeping the observed LD between the studied locus.

The power of another phylogeny-based association test has also been investigated by Seltman *et al*[19] who extended the Templeton's method[12] to case-parent trios data. But they use a simulation process different from the one used here and in Durrant *et al*[1]. Indeed, rather than choosing a susceptibility site, Seltman *et al*[19] choose groups of at-risk haplotypes on the tree and assume that all the other haplotypes are only carried by control individuals. This may give an advantage to the evolutionary-based method, especially when the at-risk haplotypes are all grouped in a clade

To conclude, the distance-based grouping of haplotypes as described in Durrant *et al*[1] will usually result in a lack in power as compared to other association tests except for very particular patterns of LD and DS frequency where a slight gain in power may be obtained. However, as suggested in Durrant *et al*[1], in Seltman *et al*[19] and in Bardel *et al*[20], the various clustering or phylogeny-based methods might be more powerful when more than one susceptibility site are involved in the disease and be more efficient to precisely localize these susceptibility sites along the haplotypes. Further investigations should confirm these points.

Supplementary information is available at European Journal of Human Genetics' web-site

Acknowledgements

We wish to thank Jean-Pierre Hugot, Habib Zouali and Suzanne Lepage from the fondation Jean Dausset for providing us with Crohn data and Caroline Durrant for kindly providing us with the software CLADHC. We also want to thank the Runtime team from the LaBRI for letting us run a part of our simulations on their cluster and two anonymous reviewers for their helpful comments.

References

1 Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P and Morris AP. Linkage disequilibrium mapping via cladistic analysis of single-nucleotique polymorphism haplotypes. *Am J Hum Genet* 2004; **75**:35–43.

2 Akey J, Jin L and Xiong M. Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? *European J of Hum Genet* 2001; **9**:291–300.

3 Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ and Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002; **53**:79–91.

4 Zhang K, Calabrese P, Nordborg M and Sun F. Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet* 2002; **71**:1386–1394.

5 Long AD and Langley CH. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999; **9**:720–731.

6 Kaplan N and Morris R. Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. *Genet Epidemiol* 2001; **20**:432–457.

7 Roeder K, Bacanu SA, Sonpar V, Zhang X and Devlin B. Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol* 2005; **28**:207–219.

8 Judson R and Stephens JC. Notes from the SNP vs. haplotype front. *Pharmacogenomics* 2001; **2**:7–10.

9 Bader JS. The relalative power of SNPS and haplotype as genetic markers for association tests. *Pharmacogenomics* 2001; **2**:11–24.

10 Culverhouse R, Suarez BK, Lin J and Reich T. A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* 2002; **70**:461–471.

11 Jannot AS, Essioux L, Reese M and Clerget-Darpoux F. Improved use of SNP information to detect the role of genes. *Genet Epidemiol* 2003; **25**:158–67.

12 Templeton AR, Boerwinkle E and Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics* 1987; **117**:343–351.

13 SeattleSNPs. NHLBI Program for Genomic Applications, UW-FHCRC, Seattle, WA. (URL: http://pga.gs.washington.edu) [accessed october 2004].

14 Stephens M, Smith NJ and Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**:978–989.

15 Stephens M and Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003; **73**:1162–1169.

16 Hugot JP, Chamaillard M, Zouali H et al. Association of NOD2 leucin-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411**:599–603.

17 Daly MJ, Kruglyak L, Pratt S et al. Genehunter, version 2.1, 2001.

18 Abecasis GR and Cookson WO. GOLD–graphical overview of linkage disequilibrium. *Bioinformatics* 2000; **16**:182–183.

19 Seltman H, Roeder K and Devlin B. Transmission/Disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet* 2001; **68**:1250–1263.

20 Bardel C, Danjean V, Hugot JP, Darlu P and Génin E. On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genetics* 2005; **6**.

GRR [a]		Single locus	T[h]	T[MAX]
Hom	Het	Power with susceptibility site (%) [b]		
2	1.5	**81.3**	46	39.6
5	1.5	**99.8**	94.2*	93.6*
10	1.5	**100***	**100***	**100***
Hom	Het	Power without susceptibility site (%) [b]		
2	1.5	**13.7***	8.6	**11.8***
5	1.5	**52.9**	31.1	38.3
10	1.5	**96.5***	87.7	**94.8***

The maximum LD between the susceptibility site (SNP 13) and another site equals 0.23. The frequency of the susceptibility allele equals 0.11.

[a] Genetic Relative Risks. Hom=homozygote, Het=heterozygote. The penetrance for homozygotes carriers of 0 disease susceptibility allele was set to 0.01.

[b] Powers (%) were evaluated on 1000 simulations. Type I error = 1%. T[MAX] refers to as the grouping method proposed by Durrant *et al* [1] and T[h] to the haplotype test without prior grouping. For each penetrance vector, the best power is indicated in bold.

* indicates that there is no statistical difference between the values (at the 5% level).

Table 1: Power of the three tests obtained over 1,000 simulations for different penetrance vectors in the CARD 15 region. The chosen susceptibility site is SNP 13

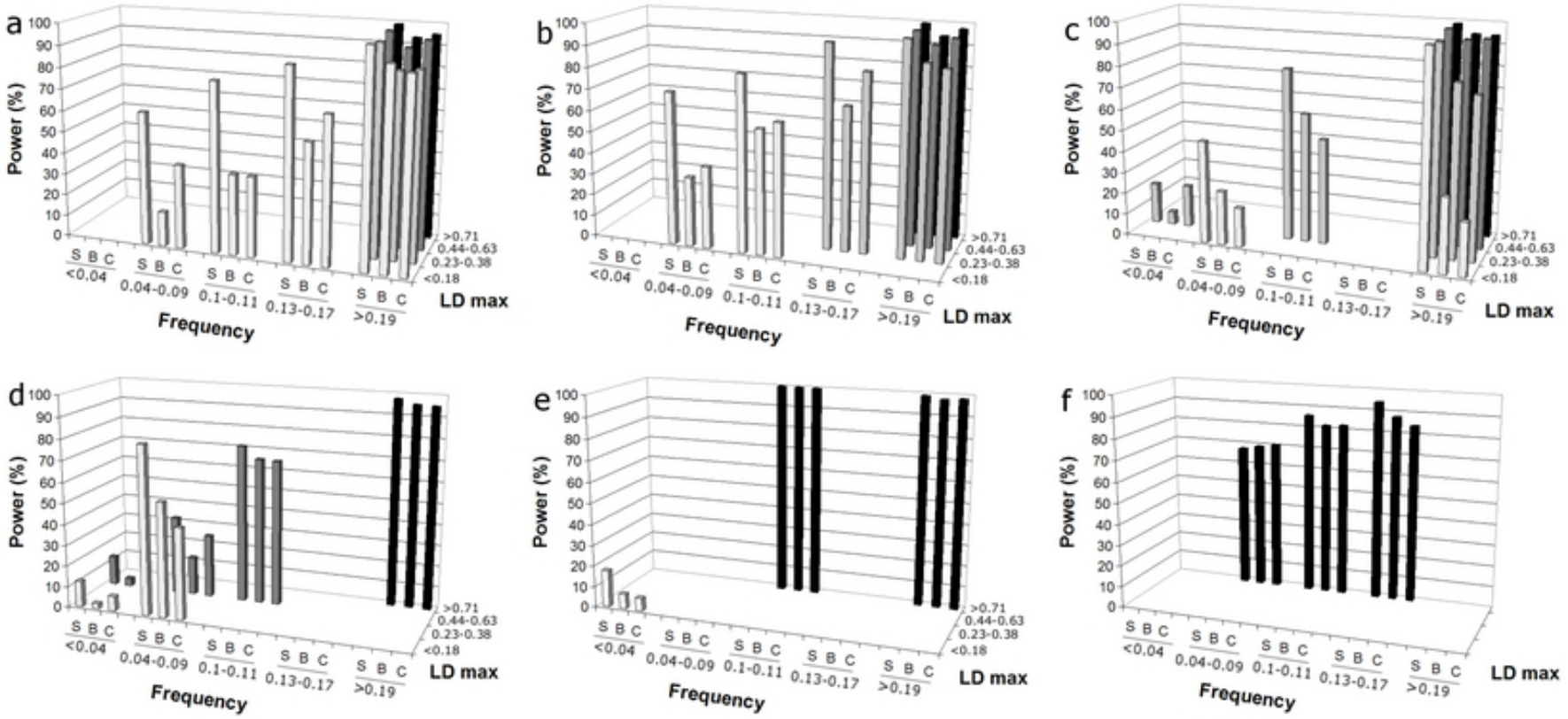


Figure 1: Power of the three tests (S: single locus test, B: T[h] test, C: T[MAX]) according to the frequency of the DS allele and to the LD_{max} between the DS site and another site. The DS site is kept in the analysis.
a: CHR20_1, b: CHR20_2, c: CARD15, d: IL13, e: PLAU, f: TNF.

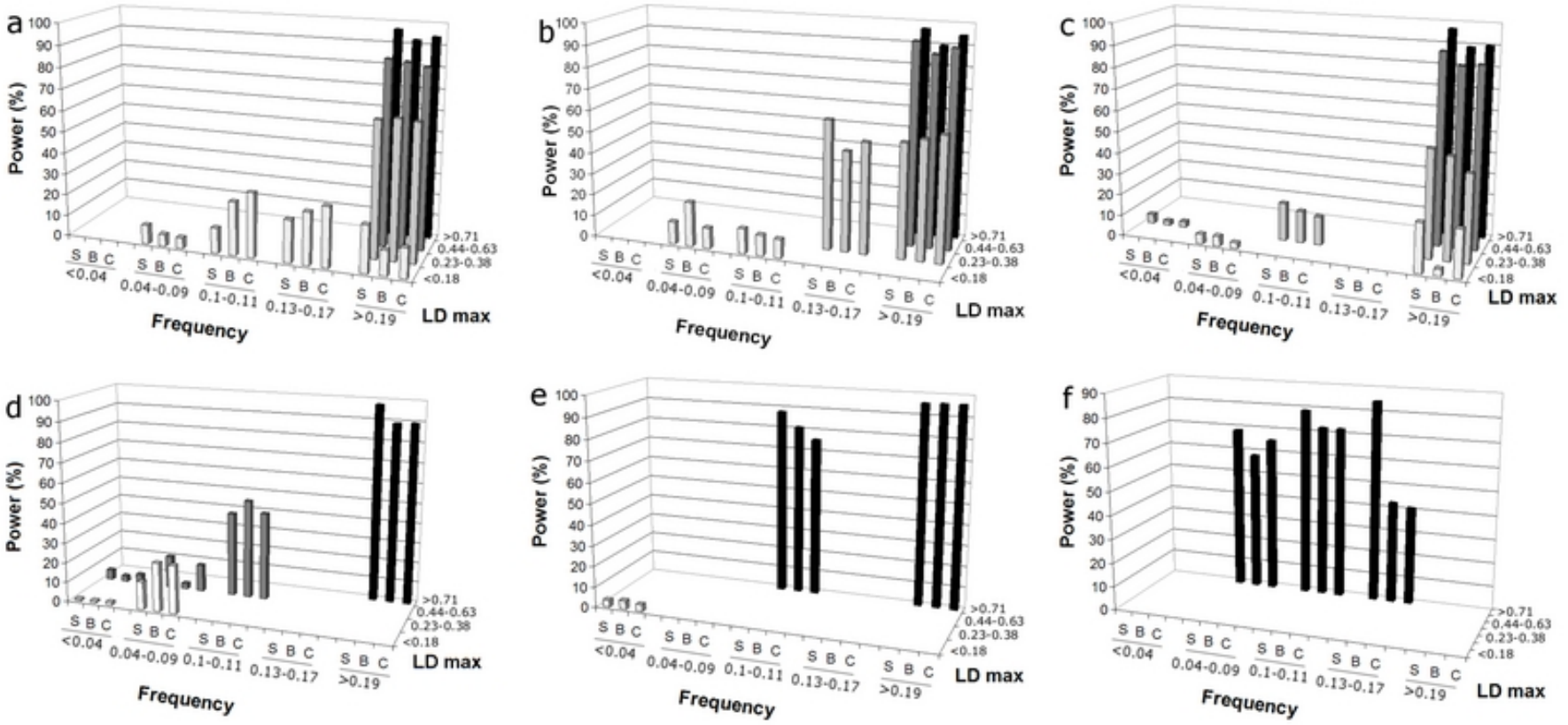


Figure 2: Power of the three tests (S: single locus test, B: T[h] test, C: T[MAX]) according to the frequency of the DS allele and to the LD_{max} between the DS site and another site. The DS site is removed before analysis.
a: CHR20_1, b: CHR20_2, c: CARD15, d: IL13, e: PLAU, f: TNF.

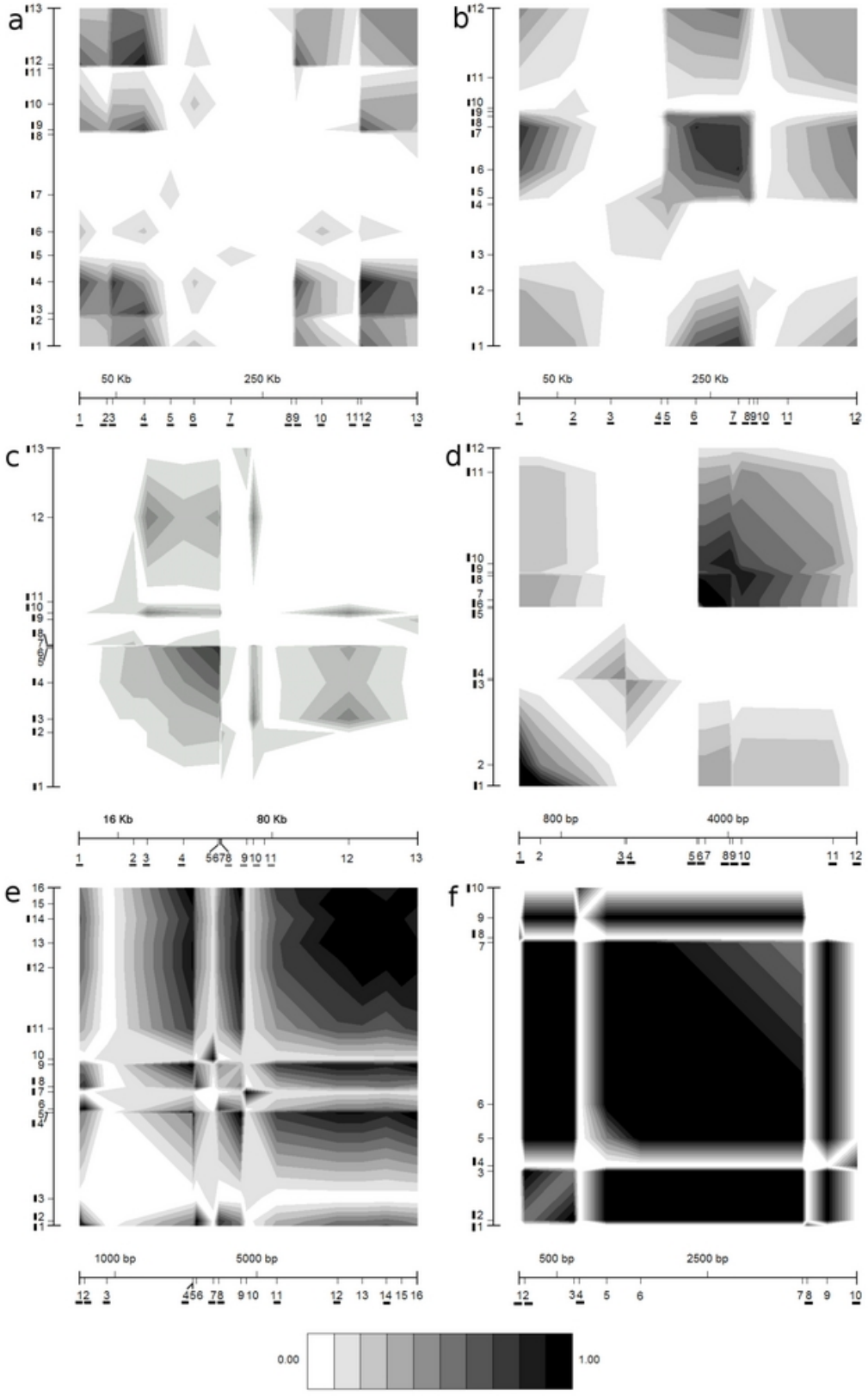


Figure 3: **(Supplementary figure)** Pairwise linkage disequilibrium for the six data sets. The LD is measured with the r^2 statistics. The sites chosen as susceptibility sites in the analysis are underlined.
a: CHR20_1, b: CHR20_2, c: CARD15, d: IL13, e: PLAU, f: TNF.

Chapitre 4

Application de la méthode à des données

Sommaire

4.1 Les données simulées pour GAW12 122
4.1.1 Description des données 122
4.1.2 Analyse des données par ALTree 122
4.1.2.1 Reconstruction de l'arbre par une méthode de parcimonie . 122
4.1.2.2 Reconstruction de l'arbre par une méthode de maximum de vraisemblance 126
4.1.3 Conclusion 126
4.2 La maladie de Crohn 128
4.2.1 Le rôle du gène CARD15 128
4.2.2 Le jeu de données 128
4.2.2.1 Obtention des données 128
4.2.2.2 Analyse descriptive des données 129
4.2.3 Analyse des données par ALTree 129
4.2.3.1 Reconstruction de l'arbre par parcimonie 130
4.2.3.2 Reconstruction de l'arbre par maximum de vraisemblance . 131
4.2.4 Conclusion 133
4.3 La schizophrénie 133
4.3.1 Le gène DRD2 : un bon gène candidat 133
4.3.2 Le jeu de données 134
4.3.2.1 Obtention des données 134
4.3.2.2 Analyse descriptive des données 134
4.3.3 Analyse des données par ALTree 136
4.3.3.1 Reconstruction de l'arbre par parcimonie 136
4.3.3.2 Reconstruction de l'arbre par maximum de vraisemblance . 137
4.3.4 Conclusion 138

Après avoir étudié la puissance d'ALTree par simulations, nous avons cherché à l'appliquer sur des données différentes de celles que nous utilisions en simulations. Nous l'avons tout d'abord appliquée à des données simulées pour un atelier de travail international (GAW 12). Dans un second temps, grâce à des collaborations avec des médecins, nous avons pu utiliser ALTree sur des données sur la maladie de Crohn et sur la schizophrénie.

4.1 Les données simulées pour GAW12

Nous avons appliqué notre méthode aux données simulées pour un atelier de travail international (*Genetic analysis Workshop 12*). Ces données sont simulées selon un processus décrit dans les articles de Thomas *et al.* (2001) et Almasy *et al.* (2001). Elles correspondent aux génotypes et aux phénotypes de 1497 individus répartis sur 23 généalogies, issus d'une population affectée par une maladie multifactorielle fréquente (prévalence[1] 25%). Cette maladie est déterminée par des facteurs génétiques et environnementaux qui influencent soit directement la probabilité pour les individus d'être malades, soit des traits quantitatifs qui vont à leurs tour modifier cette probabilité.

4.1.1 Description des données

Nous avons travaillé sur un des gènes de susceptibilité à la maladie, le gène majeur 6 (MG6), qui influence directement la probabilité pour un individu d'être atteint par la maladie. Dans notre analyse, nous nous sommes intéressés à un seul réplicat (réplicat numéro 5). Afin de travailler dans un contexte cas/témoins, nous n'avons retenu dans notre échantillon que les individus fondateurs (i.e. qui n'ont pas de parents dans la généalogie). Les haplotypes de ces individus ont été reconstruits en utilisant le logiciel `GENEHUNTER` (Daly *et al.*, 2001a)[2] après avoir découpé les généalogies en familles nucléaires constituées de l'individu fondateur dont on veut reconstruire les haplotypes, son conjoint et ses enfants. Les individus pour lesquels la reconstruction n'a pas pu se faire sans ambiguité n'ont pas été inclus dans l'échantillon. Au final, notre échantillon est constitué de 178 haplotypes (66 haplotypes de malades, 112 haplotypes de témoins), répartis en 72 catégories haplotypiques. Chaque haplotype est formé de 181 SNPs bialléliques dont 87 sont monomorphes.

4.1.2 Analyse des données par `ALTree`

4.1.2.1 Reconstruction de l'arbre par une méthode de parcimonie

Nous avons reconstruit la phylogénie des haplotypes en utilisant la méthode de parcimonie implémentée dans le logiciel `PAUP` . Les paramètres sont les mêmes que ceux décrits page 96. Nous avons obtenu 208 arbres équiparcimonieux non enracinés dont l'indice de cohérence CI vaut 0,65 (0,55 si les sites non informatifs sont éliminés). Comme nous ne disposions pas de la séquence ancestrale pour enraciner l'arbre, nous avons utilisé deux séquences ancestrales hypothétiques :

mfH : l'haplotype le plus fréquent dans l'échantillon ;

consH : une séquence consensus formée par l'ensemble des allèles les plus fréquents pour chaque SNP. Cette séquence ne faisait pas partie des haplotypes présents dans l'échantillon.

Puis, nous avons comparé les résultats obtenus avec chacune d'entre elles.

[1] La prévalence correspond à la fréquence de la maladie dans une population donnée, à un moment donné

[2] Nous avons vu au chapitre1, que le logiciel `GENEHUNTER` n'est pas très adapté pour la reconstruction haplotypique de données en présence de déséquilibre de liaison. Nous l'avons cependant utilisé car, avant mai 2004, il n'existait pas d'autres logiciels de reconstruction haplotypique pour des données familiales

Test d'association Le test d'association a été réalisé sur l'un des 208 arbres obtenus avec PAUP . Nous avons évalué la p-value en effectuant 10 000 répétitions et nous avons trouvé que, selon la séquence ancestrale choisie, le test est soit très significatif ($p_{mfH} = 7,99 \times 10^{-4}$), soit non significatif (p_{consH} de 0,07) mais tout de même relativement proche de la limite de significativité. Ces résultats contradictoires sont expliqués par le phénomène représenté par un exemple sur la Figure 4.1. Dans les deux cas, on observe un groupe constitué d'haplotypes portés préférentiellement par les malades, mais, selon l'ancêtre choisi, ce groupe ne sera pas situé au même niveau dans l'arbre.

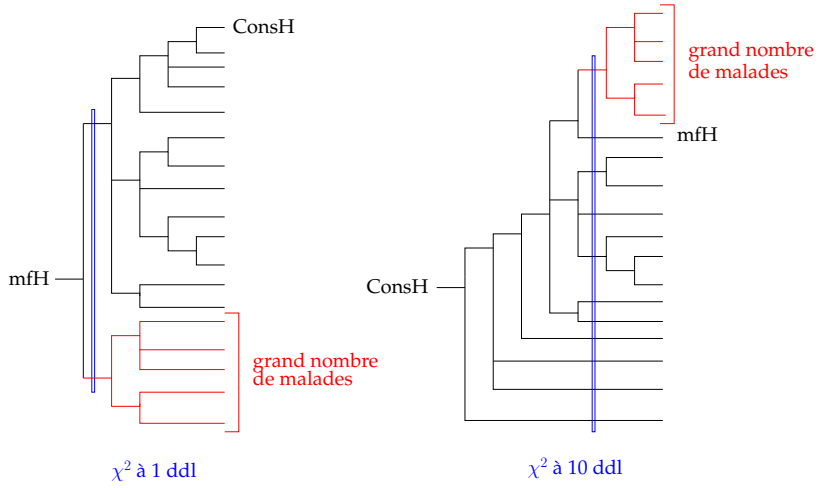


FIG. 4.1: Influence du choix de l'ancêtre sur la puissance du test d'association

Lorsque l'arbre est enraciné en utilisant l'haplotype majoritaire, le groupe contenant majoritairement des haplotypes portés par des malades est situé au 1er niveau de l'analyse. Le nombre de malades et de témoins dans ce groupe est donc comparé au nombre de malades et de témoins dans un seul autre groupe. Le test n'a qu'un seul degré de liberté et il est très significatif. Lorsque l'arbre est enraciné en utilisant l'haplotype consensus, le groupe contenant majoritairement des haplotypes portés par des malades est situé au 5e niveau sur la figure. Le test réalisé à ce niveau a alors 10 degrés de liberté et n'est pas significatif.

Dans le cas des données GAW, lorsque l'arbre est enraciné sur la séquence majoritaire, le test réalisé au niveau où le groupe contenant un excès de malades peut être détecté possède 2 degrés de liberté. Lorsque l'arbre est enraciné sur la séquence consensus, ce test possède 38 degrés de liberté. Cela explique la différence observée entre les p-values pour le test d'association.

Localisation du site de susceptibilité Le test de localisation des sites de susceptibilité a été réalisé sur les 208 arbres équiparcimonieux identifiés par PAUP avec les trois options suivantes pour l'enracinement : non enraciné, enraciné en utilisant l'haplotype le plus fréquent comme ancêtre ou enraciné en utilisant une séquence consensus comme ancêtre. Comme les résultats sont très similaires pour les trois options (V_i de même ordre de grandeur pour les différents sites), seuls les résultats obtenus avec l'arbre non enraciné sont présentés dans le

Tableau 4.1.

Pour ces analyses, nous avons choisi d'attribuer l'état " ?" (donnée indéterminée) au caractère S pour les haplotypes appartenant à des catégories représentées par un seul haplotype de malade ou un seul haplotype de témoin. De plus, les effectifs de chaque catégorie haplotypique étant faibles, le caractère S est déterminé (état différent de " ?") pour seulement 15 catégories haplotypiques sur les 72 quand le paramètre ϵ[3] est fixé à 1. Nous avons donc choisi de baisser le paramètre ϵ à 0,5 pour augmenter le nombre de catégories haplotypiques pour lesquelles le caractère S est indéterminé (ce nombre de catégories passe alors à 21). Enfin, lorsque plusieurs sites sont en déséquilibre complet ($r^2 = 1$), un seul site a été choisi pour représenter l'ensemble.

`acctran`	`deltran`
27, 30 (*5,5*)	30 (*5,5*)
10, 86, 107, 153 (*3.8*)	86 (*3,8*)
24, 161, 174 (*3*)	27 (*2,7*)
190 (*1,3*)	201 (*2,1*)
201 (*1,2*)	**10**, 40, 64, 107, 153 (*1,8*)

TAB. 4.1: Localisation du site de susceptibilité sur les 208 arbres reconstruits par parcimonie (arbre non enraciné)

L'analyse a été réalisée pour les deux méthodes d'optimisation des caractères sur l'arbre (`acctran` et `deltran`).

Pour chacune d'entre elles, nous avons retenu les sites auxquels étaient attribuées les 5 valeurs de V_i les plus élevées. Les numéros de ces sites sont indiqués dans le tableau et les valeurs des V_i correspondantes sont précisées entre parenthèses en italique.

Le site de susceptibilité simulé est indiqué en gras.

Les sites de susceptibilité les plus probables d'après ce tableau sont les sites 30, 27 et 86 puis, les sites 10, 107 et 153. Or, le site de susceptibilité simulé est le site 10. Notre méthode nous permet donc d'identifier le site de susceptibilité parmi les 6 locus les plus probables mais pas parmi les 3 locus les plus probables. Quelques observations schématisées sur la Figure 4.2 permettent d'expliquer ce résultat :

Le nombre d'individus analysés dans cette étude est relativement faible (66 malades et 112 témoins). On peut penser qu'en augmentant la taille de l'échantillon, on améliorerait la détection du site 10. En effet, la détection des sites 30 et 86 est liée au fait que ces catégories sont rares. On peut penser qu'avec un effectif plus élevé, les branches sur lesquelles ces sites changent ne seraient pas détectées. On peut aussi espérer qu'une augmentation de l'effectif permette d'identifier la deuxième branche sur laquelle le site 10 change d'état.

On note qu'en utilisant d'autres paramètres pour définir les états du caractère S pour les différents haplotypes et, en particulier, en utilisant l'information des catégories haplotypiques représentées par un seul haplotype, on obtient des résultats légèrement différents. Le V_i du site 10 est le deuxième plus élevé derrière celui du site 221 lorsqu'`acctran` est utilisé. Avec `deltran`, le V_i du site 10 est le troisième derrière celui du site 221 et du groupe de sites 30, 109, 189 et 219.

[3]Pour une description du paramètre ϵ, voir page 66

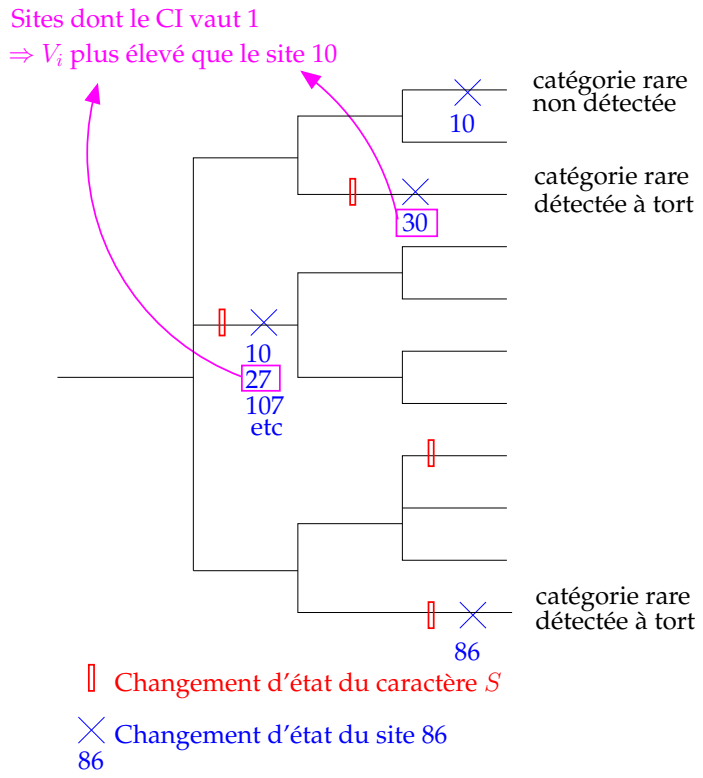


FIG. 4.2: Exemple d'arbre illustrant les problèmes observés pour la localisation du site de susceptibilité pour les données GAW

Le site 10 change d'état sur deux branches de l'arbre, dont une branche terminale menant à une catégorie haplotypique représentée par un seul malade. Or, pour les catégories haplotypiques ne comportant qu'un seul élément, nous avons choisi d'attribuer l'état indéterminé au caractère S. Notre programme ne peut donc détecter le site 10 que sur l'autre branche.

Sur cette autre branche, 11 autres sites changent d'état dont, en particulier, le site 27 qui ne change d'état que sur cette branche. Cela explique que le V_i de ce site soit supérieur à celui du site 10 qui n'est détecté qu'une seule fois, alors qu'il change deux fois d'état sur l'arbre.

En plus de cette branche, le caractère S change d'état sur 3 autres branches. Ces trois branches sont des branches terminales dans lesquelles l'effectif est faible. Elles sont systématiquement détectée. Or, sur deux de ces branches, les sites 86 et 30 changent d'état. Le site 30 ne mutant que sur cette branche et cette branche étant systématiquement détectées, cela explique qu'il ait le plus fort V_i.

4.1.2.2 Reconstruction de l'arbre par une méthode de maximum de vraisemblance

Nous avons reconstruit la phylogénie en utilisant le logiciel PHYML et nous avons reconstruit les états de caractère en utilisant le logiciel PAML . Nous avons utilisé les mêmes paramètres de reconstruction de l'arbre que ceux décrits page 74, à l'exception de la proportion de sites invariants que nous avons estimée au lieu de la fixer à zéro car il y a de nombreux sites invariants dans les données GAW.

Test d'association Le test d'association a été réalisé sur l'arbre le plus vraisemblable. Comme pour l'analyse par parcimonie, nous obtenons des p-values très différentes selon la séquence ancêtre choisie. Lorsque l'ancêtre est l'haplotype majoritaire, le test est très significatif ($p_{mfH} = 5 \times 10^{-4}$) alors qu'il ne l'est pas du tout lorsque l'ancêtre est l'haplotype consensus ($p_{consH} = 0.7$). Il se produit un phénomène semblable à celui décrit sur la Figure 4.1 dans le cas où la méthode de parcimonie est utilisée. Les nombres de degrés de liberté des tests effectués au niveau de l'arbre où se situe le groupe contenant une majorité d'haplotypes malades sont encore plus différents avec la méthode de maximum de vraisemblance qu'avec la méthode de parcimonie (2 ddl avec mfH, 56 ddl avec consH).

Localisation du site de susceptibilité Le test de localisation a été réalisé sur l'arbre non enraciné le plus vraisemblable. Comme pour l'analyse de parcimonie, nous avons choisi d'attribuer l'état " ?" au caractère S pour les catégories haplotypiques représentées par un seul haplotype et de fixer la valeur de ϵ à 0,5. Les résultats sont récapitulés dans le Tableau 4.2.

sites	V_i
27 et 30	5,5
13	4.7
10, 64, 86 et 153	3.7
17	3,4
24 et 12	2,9

Tous les sites dont le V_i est supérieur à zéro sont présentés dans le tableau.

TAB. 4.2: Localisation du site de susceptibilité sur l'arbre de vraisemblance maximum (arbre non enraciné)

Les résultats sont assez semblables à ceux obtenus sur l'arbre reconstruit par parcimonie. Le site de susceptibilité n'est pas l'un des deux sites pour lesquels le V_i est maximum mais il est détecté parmi les sites possédant les 3 valeurs maximales de V_i. Comme pour les arbres reconstruits par parcimonie, plusieurs sites (7) changent d'état sur la même branche que le site de susceptibilité.

4.1.3 Conclusion

Cette étude illustre bien certains problèmes de notre méthode. Tout d'abord, concernant la détection de l'association, le niveau de l'arbre où est détecté le groupe d'haplotypes essentiellement porté par des malades va énormément influencer la puissance de détection de l'association. Or, ce niveau dépend beaucoup de la position de la racine de l'arbre. Le

choix de la séquence ancestrale est donc très important. Dans l'exemple des données GAW, on remarque que l'enracinement sur l'haplotype majoritaire donne un meilleur résultat que l'enracinement sur l'haplotype consensus. Cependant, dans l'absolu, il est assez difficile de choisir entre les deux méthodes d'enracinement. D'après la théorie de la coalescence, l'haplotype majoritaire doit être l'haplotype ancêtre. Il possède l'avantage d'être toujours présent dans l'échantillon, ce qui permet d'être certain qu'il s'agit d'un haplotype existant réellement dans la population. Cependant, comme les individus portant l'haplotype ancêtre ne vont pas être intégrés dans l'analyse, le fait de choisir l'haplotype majoritaire peut parfois faire beaucoup baisser la taille de l'échantillon. C'est pourquoi il est conseillé de tester plusieurs méthodes d'enracinement.

Il faut cependant souligner que le problème de la position du groupe d'haplotypes portés essentiellement par des individus malades se pose surtout pour les jeux de données constitués d'un grand nombre de catégories haplotypiques, puisque le nombre de degrés de liberté du test en dépend.

Concernant la localisation du site de susceptibilité, on remarque que lorsque plusieurs sites mutent sur la même branche que le site de susceptibilité à la maladie, le V_i de tous ces sites est élevé, ce qui rend plus difficile l'identification précise du site de susceptibilité. Le second problème mis en évidence par cet exemple est celui des faibles effectifs. En effet, lorsqu'une catégorie haplotypique n'est représentée que par un faible nombre d'haplotypes, et en particulier, par un seul haplotype, il est difficile de lui attribuer l'état du caractère "S". Nous avons testé deux possibilités :

- soit imposer l'état du caractère en fonction du statut du porteur de l'haplotype. Mais, le risque d'erreur dans l'attribution de l'état du caractère S peut être assez élevé si un individu hétérozygote porte un haplotype rare, ou dans le cas de phénocopies ou de pénétrance incomplète. C'est ce qui nous fait détecter à tort le site 221 lors de l'analyse des arbres reconstruits par la méthode de parcimonie[4] ;
- soit imposer l'état "indéterminé" à ces catégories. Dans ce cas, il ne sera pas possible de détecter des changements du site de susceptibilité sur des branches menant à une catégorie rare. Le V_i du site de susceptibilité risque alors d'être plus faible que celui d'autres sites, en particulier, de sites qui ne changent d'état que dans une seule branche de l'arbre sur laquelle un changement d'état de S est observé.

Il existe une troisième solution qui consiste à éliminer les haplotypes rares de l'échantillon. Cette solution est relativement souvent employée dans la littérature (par exemple Zhu *et al.* (2000); Rodriguez *et al.* (2004); Keavney *et al.* (1998), etc.). Cela implique une certaine perte d'information. Sur les données GAW, nous avons éliminé les catégories haplotypiques ne contenant qu'un seul haplotype. Il nous reste alors 38 catégories haplotypiques que nous avons utilisées pour reconstruire un arbre par la méthode de maximum de vraisemblance. L'analyse de l'arbre le plus vraisemblable nous a permis cette fois de détecter le site de susceptibilité parmi les 4 sites ayant le V_i le plus élevé (V_i=4,8).

Il est donc probable qu'en fonction des jeux de données, l'un ou l'autre de ces traitements des catégories rares soit plus efficace. Il est alors intéressant de réaliser les analyses dans les trois conditions, puis de les comparer.

[4] le même phénomène s'observe aussi sur l'arbre reconstruit par maximum de vraisemblance

4.2 La maladie de Crohn

La maladie de Crohn est une maladie inflammatoire de l'intestin qui apparaît la plupart du temps lors de l'adolescence (15 à 25 ans). Ses symptômes principaux sont des douleurs et des crampes abdominales, des diarrhées chroniques, une perte de poids et, parfois, de la fièvre. Actuellement, l'incidence de cette maladie tend à augmenter : inconnue avant 1920, sa prévalence atteint maintenant 0,15% dans les pays d'Europe de l'ouest et en Amérique du nord (Schreiber *et al.*, 2005).

Les causes exactes de la maladie de Crohn ne sont actuellement pas connues. L'augmentation de la prévalence suggère que des facteurs environnementaux liés au mode de vie seraient impliqués : présence d'agents infectieux, modification du régime alimentaire, etc. D'autre part, de nombreux arguments montrent le rôle de facteurs génétiques :

- forte agrégation familiale : le risque de la maladie chez les frères et sœurs d'un individu atteint de la maladie de Crohn est 15 à 35 fois supérieur au risque de la population générale (Schreiber *et al.*, 2005) ;
- taux de concordance de 20 à 50% chez les jumeaux monozygotes alors qu'il est de 0 à 4% chez les dizygotes (Brant et Shugart, 2004).

La maladie de Crohn a donc une étiologie complexe et il est maintenant admis qu'il existe non pas un gène responsable, mais plusieurs facteurs génétiques de risque, dont l'action peut être modifiée par des facteurs environnementaux.

4.2.1 Le rôle du gène CARD15

La première région impliquée dans la susceptibilité à la maladie de Crohn a été découverte par analyse de liaison sur l'ensemble du génome (Hugot *et al.*, 1996). Il s'agit de la région péri-centromérique du chromosome 16 qui a été nommée IBD1 (Inflammatory Bowel Disease locus 1) par Hugot *et al.*. Cette découverte a été répliquée par de nombreuses analyses de liaison ; c'est pourquoi, à la fin des années 90, cette région a fait l'objet de nombreuses études. Finalement, en 2001, deux équipes ont simultanément montré l'implication du gène CARD15 (codant pour la protéine NOD2) dans la susceptibilité à la maladie de Crohn (Hugot *et al.*, 2001; Ogura *et al.*, 2001) et ont identifié des polymorphismes associés à la maladie : R702W (SNP 8) et G908R (SNP 12) (Hugot *et al.*, 2001) et 1007fsInsC (SNP13) (Hugot *et al.*, 2001; Ogura *et al.*, 2001).

Nous avons cherché à répliquer ces résultats en utilisant notre méthode phylogénétique.

4.2.2 Le jeu de données

4.2.2.1 Obtention des données

Les données nous ont été fournies par Jean-Pierre Hugot. Elles sont constituées de 235 familles avec au moins un enfant atteint de la maladie de Crohn. Tous les individus ont été génotypés pour 13 SNPs[5] dans le gène CARD15. La Figure 4.3 représente la position des 13 marqueurs dans la région CARD15.

Les haplotypes de tous les individus ont été reconstruits par `GENEHUNTER 2.0b` (Daly *et al.*, 2001a). Nous avons ensuite conservé un seul enfant atteint par famille, choisi aléatoirement, et nous avons éliminé les haplotypes où il manquait des données pour plus de 2 mar-

[5]Pour une description plus complète de la méthode de séquençage, voir Hugot *et al.* (2001)

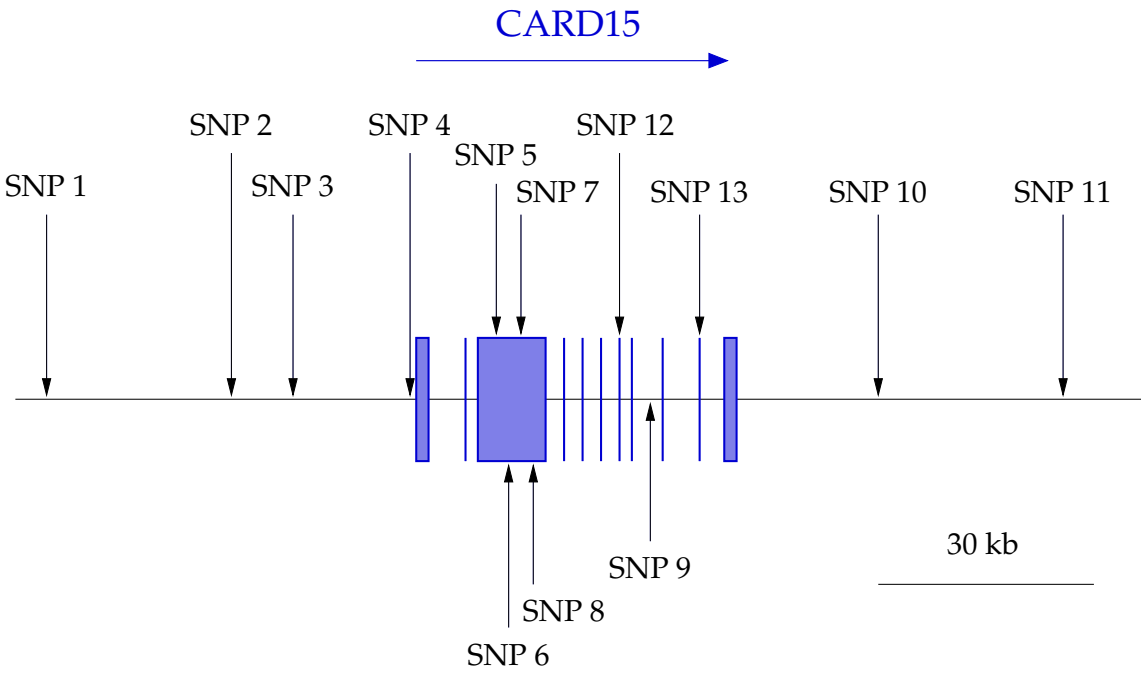


FIG. 4.3: Organisation de la région contenant le gène CARD15 et localisation des 13 polymorphismes

queurs. Nous avons obtenu un échantillon de 375 haplotypes de malades. Les haplotypes des parents non transmis à l'enfant atteint sélectionné ont alors été affectés à l'échantillon de témoins. Après élimination des haplotypes pour lesquels il manquait des données pour plus de deux marqueurs, cet échantillon est finalement constitué de 351 haplotypes. Au total, les 726 haplotypes de l'échantillon global se répartissent en 126 catégories haplotypiques.

4.2.2.2 Analyse descriptive des données

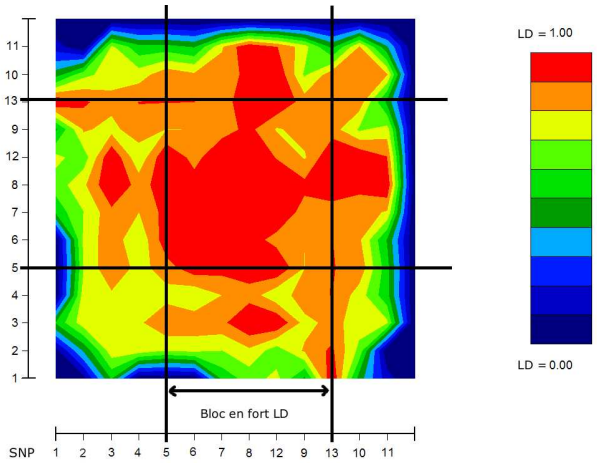
Les fréquences alléliques des différents SNPs sont données sur la Figure 4.4a et la matrice de déséquilibre de liaison (produite par le logiciel `GOLD` (Abecasis et Cookson, 2000)) est représentée sur la Figure 4.4b. On remarque que les marqueurs situés entre le SNP 5 et le SNP 13 forment un bloc en fort déséquilibre de liaison.

4.2.3 Analyse des données par `ALTree`

Nous avons reconstruit l'arbre racontant l'histoire évolutive des 126 haplotypes en utilisant la méthode de parcimonie implémentée dans le logiciel `PAUP` . Nous avons retenu 1000 arbres équiparcimonieux (1000 étant la limite que nous avons fixée à `PAUP`), dont le CI vaut 0,0909. La quantité d'homoplasie est donc très importante dans ce jeu de données et nous pouvons penser qu'elle est en partie due à de la recombinaison. Nous avons donc décidé de limiter l'analyse à une région du gène CARD15 dont on pouvait supposer qu'elle contenait peu de recombinaisons. Pour cela, nous nous sommes intéressés au déséquilibre de liaison entre les marqueurs. La Figure 4.4b suggère que les SNPs 5 à 13 forment un bloc. Cette conclusion correspond aux résultats de Vermeire *et al.* (2002) obtenus sur une popula-

	témoins	malades
marqueurs	allèle 1	allèle 1
SNP 1	0,37	0,40
SNP 2	0,77	0,62
SNP 3	0,64	0,65
SNP 4	0,61	0,50
SNP 5	0,39	0,50
SNP 6	0,39	0,50
SNP 7	0,64	0,68
SNP 8	0,10	0,16
SNP 12	0,97	0,94
SNP 9	0,40	0,32
SNP 13	0,95	0,87
SNP 10	0,59	0,67
SNP 11	0,89	0,89

(a) Fréquences alléliques des 13 SNPs chez les témoins et chez les malades



Les 13 SNPs sont représentés en abcisse et en ordonnée. Dans la matrice, les valeurs de déséquilibre de liaison (LD) pour chaque paires de SNPs sont représentée selon le code de couleur expliqué sur l'échelle, à droite de la matrice. Le LD de la paire de SNPs (A,B) étant le même que celui de la paire (B,A), la matrice est symétrique.

(b) Matrice de LD (mesure de déséquilibre : D')

FIG. 4.4: Analyse descriptive des données sur la maladie de Crohn

tion canadienne. C'est pourquoi, dans la suite de notre analyse, nous ne travaillerons plus que sur ces 7 SNPs, qui définissent 33 haplotypes différents. Par ce processus, nous obtenons un nouveau jeu de données contenant beaucoup moins d'homoplasies : le CI des arbres reconstruits par parcimonie vaut alors 0.32.

4.2.3.1 Reconstruction de l'arbre par parcimonie

Nous avons reconstruit la phylogénie en utilisant la méthode de parcimonie implémentée dans PAUP [6]. Comme nous ne disposions ni de la séquence ancestrale ni de séquences de singes que l'on aurait pu utiliser comme groupe externe pour enraciner l'arbre, nous avons utilisé deux séquences ancestrales hypothétiques : l'haplotype majoritaire (mfH) ou une séquence consensus formée des allèles les plus fréquents à tous les locus (consH). Contrairement à ce que l'on avait observé pour les données GAW, la séquence consensus est un des haplotypes du jeu de données. Il s'agit du second haplotype le plus fréquent dans l'échantillon.

Test d'association Le test d'association a été réalisé sur le premier arbre obtenu par PAUP . Nous avons évalué la p-value sur 100 000 répétitions et nous avons trouvé que, quelle que soit la séquence ancestrale choisie, le gène CARD15 est fortement associé à la maladie de Crohn : $p_{mfH} = 2 \times 10^{-5}$, $p_{consH} = 3 \times 10^{-5}$. Contrairement à ce que nous avions observé sur les données du GAW 12, il n'y a pas ici de différence entre les résultats obtenus avec les

[6] Pour plus de détails concernant les options de PAUP utilisées dans cette analyse, voir Bardel *et al.* (2005a)

deux séquences ancestrales. C'est probablement lié au fait qu'il y a peu d'haplotype dans cet échantillon (33) par rapport à celui du GAW 12 (72). Le nombre de degrés de liberté des tests varie donc moins avec la séquence ancestrale choisie.

Localisation des sites de susceptibilité Le test de localisation des sites de susceptibilité a été réalisé sur les 1000 premiers arbres retenus par PAUP . En classant les sites par ordre croissant de V_i, nous obtenons les résultats du Tableau 4.3

	V_i faible ——————————> V_i élevé						
mfH [a]	SNP 5	SNP 9	SNP 6	SNP 7	**SNP 13**	**SNP 12**	**SNP8**
consH [b]	SNP 9	SNP 5	SNP 6	SNP 7	**SNP 13**	**SNP 12**	**SNP8**

[a] Ancêtre = haplotype le plus fréquent
[b] Ancêtre = séquence consensus

TAB. 4.3: Analyse globale du jeu de données Crohn sur des arbres reconstruits par parcimonie
Les sites précédemment identifiés par Hugot *et al.* (2001) comme étant impliqués dans la maladie sont indiqués en gras.

On remarque que quelle que soit la séquence ancestrale choisie, les trois SNPs dont le V_i est le plus élevé sont les SNPs 8, 12 et 13, qui sont les trois sites identifiés par Hugot *et al.* (2001) comme étant impliqués dans le déterminisme de la maladie de Crohn.

Cette analyse ne nous permet cependant pas de définir si un ou plusieurs sites sont impliqués dans le déterminisme de la maladie. Nous avons alors réalisé une analyse pas à pas inspirée de la méthode des *Relative Predispositional Effects* de Payami *et al.* (1989) qui consiste, pour un test d'association cas/témoins réalisé sur un marqueur pluri-allélique, à éliminer l'allèle avec l'effet le plus important et à recommencer le test avec les allèles restant. Sur nos données, nous avons donc réalisé un premier test en considérant tous les SNPs et nous avons identifié celui qui a le plus fort V_i. Les haplotypes portant l'allèle de susceptibilité pour ce SNP sont alors éliminé de l'échantillon et un test d'association est réalisé sur ces nouvelles données. La p-value obtenue est appelée p-value résiduelle. Si elle est inférieure à un seuil (5%), on considère qu'il existe encore un SNP de susceptibilité dans ce jeu de données et un nouveau test de localisation est réalisé. Ce processus est répété jusqu'à ce que le test d'association ne soit plus significatif (p-value résiduelle supérieure à 5%).

Les résultats obtenus sont détaillés dans le Tableau 4.4

Selon la méthode d'enracinement, nous n'obtenons pas exactement les mêmes résultats. Lorsque l'haplotype majoritaire est utilisé comme séquence ancestrale, un SNP de plus, le SNP 9, est identifié. Cependant, lorsque les haplotypes portant l'allèle à risque de ce SNP sont éliminés de l'échantillon, l'association entre le gène et la maladie est encore plus significative que lorsque ces haplotypes à risque sont présents. Il est donc peu probable que le SNP 9 ait vraiment un rôle dans la susceptibilité à la maladie. Finalement, trois SNPs (SNP 8, SNP 12 et SNP 13) peuvent être identifiés comme étant des facteurs de risque.

4.2.3.2 Reconstruction de l'arbre par maximum de vraisemblance

Nous avons reconstruit la phylogénie en utilisant le logiciel PHYML et nous avons reconstruit les états de caractères en utilisant le logiciel PAML . Nous avons utilisé les mêmes

	mfH[a]	consH[b]
1er SNP identifié	8	8
p-value résiduelle[d]	4×10^{-5}	3×10^{-5}
2e SNP identifié	12	12
p-value résiduelle[d]	8×10^{-5}	3×10^{-5}
3e SNP identifié	9	13
p-value résiduelle[d]	2×10^{-5}	NS[c]
4e SNP identifié	13	
p-value résiduelle[d]	NS[c]	

[a] Ancêtre = haplotype le plus fréquent
[b] Ancêtre = séquence consensus
[c] Non significatif à 5%
[d] La p-value résiduelle correspond à la p-value obtenue après avoir éliminé du jeu de données les haplotypes portant l'allèle de susceptibilité

TAB. 4.4: Résultats de l'analyse détaillée du jeu de données Crohn par parcimonie

paramètres de reconstruction de l'arbre que ceux décrits page 74.

Test d'association Le test d'association a été réalisé sur l'arbre dont la vraisemblance est la plus élevée. La p-value a été évaluée par 10 000 permutations et nous avons trouvé que le gène CARD15 est fortement associé à la maladie de Crohn quelle que soit la séquence ancestrale utilisée : $p_{mfH} = 3 \times 10^{-4}; p_{consH} = 10^{-4}$. Les résultats obtenus avec les deux séquences ancestrales sont toujours de même ordre de grandeur.

Localisation des sites de susceptibilité L'arbre unique retenu par maximum de vraisemblance n'est pas enraciné et, contrairement à ce que nous avons fait pour les arbres de parcimonie, nous ne l'avons pas enraciné lors du test de la localisation. L'analyse globale du jeu de données nous donne les résultats du Tableau 4.5.

V_i faible ——————————> V_i élevé						
SNP 5	SNP 6	SNP 9	SNP 7	**SNP 8**	**SNP 12**	**SNP13**

Les sites précédemment identifiés par Hugot *et al.* (2001) comme étant impliqués dans la maladie sont indiqués en gras

TAB. 4.5: Analyse globale du jeu de données Crohn par maximum de vraisemblance

On remarque que les sites avec les plus forts V_i sont les trois mêmes que par parcimonie, mais que l'ordre des trois sites est différent. De même que par parcimonie, nous avons fait une analyse pas à pas. Dans cette analyse, il y a une étape où l'association est testée. Pour cette étape, nous avons enraciné les arbres sur les deux séquences mfH et consH. Nous avons successivement mis en évidence les SNPs suivants :

– SNP 13, p_{consH} résiduelle : 4×10^{-4} et p_{mfH} résiduelle : 1.3×10^{-3} ;
– SNP 12, p_{consH} résiduelle : 4×10^{-4} et p_{mfH} résiduelle : 1.6×10^{-3} ;
– SNP 8, p_{consH} résiduelle : $0,16$ et p_{mfH} résiduelle : $0,12$.

4.2.4 Conclusion

Quelle que soit la méthode phylogénétique employée, nous avons détecté 3 SNPs potentiellement impliqués dans le déterminisme de la maladie de Crohn situés sur le gène CARD15 : SNP 8 et SNP 12 dont les mutations correspondent à des substitutions d'acides aminés et SNP 13, dont la mutation correspond à une insertion d'une paire de bases. Il est intéressant de noter que ces 3 SNPs sont ceux qui avaient été initialement mis en évidence par Hugot *et al.* (2001). Depuis, le rôle de ces 3 polymorphismes a été confirmé dans plusieurs populations (par exemple dans une population canadienne, Vermeire *et al.* (2002)). De plus, Ogura *et al.* (2001) ont montré que la mutation du SNP 13 avait des conséquences fonctionnelles qui pouvaient être liées à la maladie de Crohn[7].

4.3 La schizophrénie

La schizophrénie, du grec "schizo", séparé et "phên" esprit, est une maladie psychiatrique fréquente (prévalence : 1%). Ses symptômes principaux sont très variés — idées délirantes, hallucinations, comportements et discours désorganisés, manque de motivation, perturbation du fonctionnement social pouvant aller jusqu'à l'isolement total, etc.— et leur expression peut être très différente d'un individu à l'autre.

Les causes de la schizophrénie ne sont pas connues précisément. Les études familiales ont montré l'existence d'une concentration familiale de la schizophrénie (10% de risque d'occurrence de la maladie dans une fratrie comparé à 1% dans la population générale). L'étude des taux de concordance chez les jumeaux monozygotes (48%) et dizygotes (17%) ainsi que des risques d'occurrence de la maladie plus élevés chez les enfants adoptés issus d'un parent schizophrène confirment l'existence de facteurs de risque génétique (pour une revue voir McGuffin *et al.* (1995)). Cependant, le taux de concordance inférieur à 50% chez les jumeaux monozygotes montre que des facteurs environnementaux sont aussi impliqués dans le déterminisme de la schizophrénie.

4.3.1 Le gène DRD2 : un bon gène candidat

Pour la recherche de facteurs génétiques de risques de la schizophrénie, différents arguments suggèrent que le gène codant pour le récepteur D2 de la dopamine est un bon gène candidat.

Arguments physio-pathologiques : (pour une revue, voir Seeman (1987))
– les médicaments employés dans le traitement de la schizophrénie (neuroleptiques) sont des antagonistes dopaminergiques [8] ;

[7] L'une des causes possibles de la maladie de Crohn est une réponse inflammatoire anormale déclenchée par la présence de micro-organismes de la flore intestinale. Or, l'insertion d'une paire de bases (mutation au SNP13) entraîne la synthèse d'une protéine NOD2 tronquée de 33 acides aminés. Ogura *et al.* ont montré que cette protéine tronquée ne permet plus d'activer le facteur de transcription NF-κB en présence de lipopolisaccharide bactérien

[8] c'est-à-dire qu'ils induisent une baisse de l'activité du système dopaminergique en bloquant ses récepteurs

- les substances qui entraînent la libération de dopamine nouvellement synthétisée, comme par exemple les amphétamines, peuvent induire *de novo* des symptômes psychotiques chez des individus sains ;
- des études post-mortem ou des études par tomographie par émission de positrons ont montré que la densité de récepteur D2 est plus élevée chez les schizophrènes que dans la population générale (Seeman, 1987; Kestler *et al.*, 2001).

Arguments génétiques : différentes méta-analyses[9] d'études d'association ont confirmé que des polymorphismes du gène DRD2 sont associés à la schizophrénie. Deux méta-analyses ont montré que l'allèle C du marqueur S311C (marqueur 4 dans notre analyse) est un facteur de risque pour la schizophrénie (Jönsson *et al.*, 2003; Glatt *et al.*, 2003) et une autre méta-analyse a montré que l'allèle A2 du polymorphisme TaqI A1/A2 (marqueur 3 dans notre analyse) est associé à la schizophrénie (Dubertret *et al.*, 2001).

4.3.2 Le jeu de données

4.3.2.1 Obtention des données

Les données ont été recueillies par C. Dubertret. Elles sont constituées de 101 patients atteints de schizophrénie selon les critères DSM-IV, (*Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition* [10]), recrutés dans un hôpital français et de leurs deux parents. Tous les individus ont été génotypés pour 7 SNPs et un microsatellite dans la région DRD2/X-kinase. La Figure 4.5 représente la position des différents marqueurs dans la région DRD2/X-kinase.

Pour une description plus complète de la méthode de génotypage, voir Dubertret *et al.* (2004). Les haplotypes des différents individus ont été reconstruits en utilisant FAMHAP v1.2 (Becker et Knapp, 2004a). A partir du jeu de données initial, nous avons obtenu 59 catégories haplotypiques.

De même que pour l'analyse des données sur la maladie de Crohn, nous avons affecté les 202 haplotypes des enfants au groupe des malades et les 202 haplotypes non transmis des parents au groupe des témoins.

4.3.2.2 Analyse descriptive des données

Les fréquences alléliques des différents SNPs sont données sur la Figure 4.6a et la matrice de déséquilibre de liaison (produite par le logiciel GOLD (Abecasis et Cookson, 2000)) est représentée sur la Figure 4.6b. Le microsatellite possède 6 états de caractères dont les fréquences sont $f(1) = 0,007$, $f(2) = 0,138$, $f(3) = 0,143$, $f(4) = 0,480$, $f(5) = 0,22$ et $f(6) = 0,005$.

[9] Une méta-analyse consiste à rassembler les résultats de plusieurs études indépendantes (souvent publiées dans la littérature) et à les ré-analyser pour en tirer une conclusion globale.

[10] Ce manuel, publié en 1994 par l'association américaine de psychiatrie regroupe les critères de diagnostique des diverses maladies mentales. Pour la schizophrénie, ces critères peuvent être consultés sur le site web suivant : http://cebmh.warne.ox.ac.uk/cebmh/elmh/nelmh/schizophrenia/diagnosis/dsm/page1.html

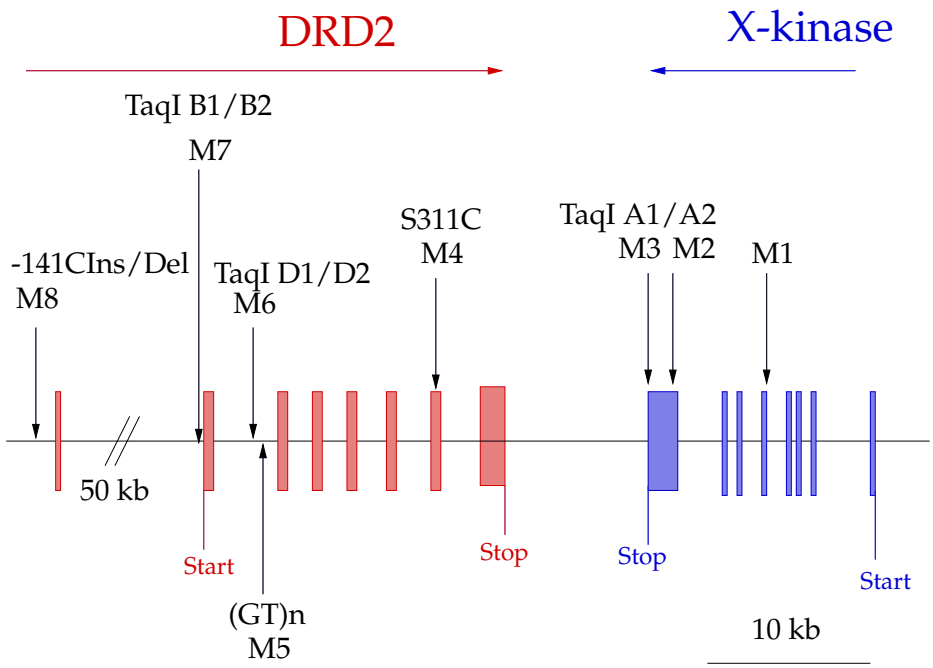
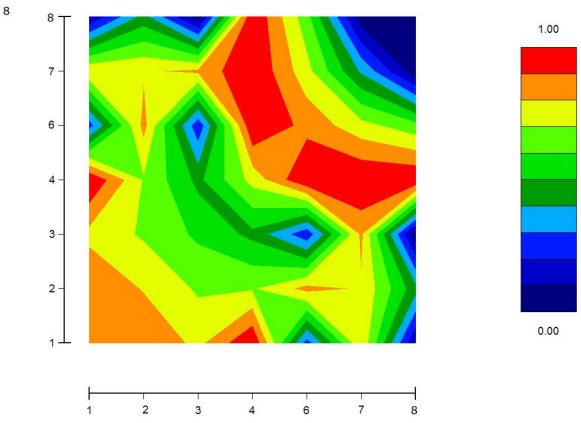


FIG. 4.5: Organisation de la région DRD2/X-kinase et localisation des 8 polymorphismes étudiés.

	témoins	malades
marqueurs	allèle 1	allèle 1
M1	0,85	0,79
M2	0,52	0,53
M3	0,16	0,27
M4	0,03	0,035
M6	0,61	0,61
M7	0,11	0,15
M8	0,10	0,13

(a) Fréquences alléliques des 7 SNPs



Les 7 SNPs sont représentés en abcisse et en ordonnée. Dans la matrice, les valeurs de déséquilibre de liaison (LD) pour chaque paires de SNPs sont représentée selon le code de couleur expliqué sur l'échelle, à droite de la matrice.

(b) Matrice de LD (mesure de déséquilibre : D')

FIG. 4.6: Analyse descriptive des données sur la schizophrénie

4.3.3 Analyse des données par ALTree

Nous avons reconstruit la phylogénie sur les 59 catégories haplotypiques. Lorsque c'était nécessaire, nous avons enraciné l'arbre sur l'haplotype consensus formé des allèles les plus fréquents à chaque locus. Cet haplotype est le 3e haplotype le plus fréquent dans l'échantillon. Nous n'avons pas utilisé l'haplotype majoritaire comme ancêtre car il est présent 123 fois dans un échantillon total de 404 haplotypes. L'haplotype utilisé comme ancêtre n'étant pas pris en compte lors de l'analyse avec ALTree , la taille de notre échantillon aurait été diminuée d'un quart si l'arbre avait été enraciné sur l'haplotype majoritaire.

4.3.3.1 Reconstruction de l'arbre par parcimonie

Nous avons utilisé le logiciel PAUP avec les paramètres décrits page 73. Nous avons choisi de conserver le microsatellite pour reconstruire l'arbre car ce type de caractère est très informatif pour la reconstruction phylogénétique. En effet, nous avons considéré les états de caractères comme ordonnés, c'est-à-dire que seul le passage d'un état à l'état voisin est possible ($1 \longleftrightarrow 2 \longleftrightarrow 3 \longleftrightarrow 4 \longleftrightarrow 5 \longleftrightarrow 6$), ce qui apporte beaucoup de contraintes pour la reconstruction de l'arbre. De plus, nous avons vérifié que sa présence n'augmente pas beaucoup l'homoplasie (Le CI passe de 0,25 lorsque le microsatellite est éliminé du jeu de données à 0,21 lorsqu'il est présent). En revanche, comme il possède plusieurs états, la probabilité qu'il change d'état sur la même branche que le caractère S du simple fait du hasard est relativement élevée. Nous avons donc choisi de ne pas le prendre en compte au cours de l'analyse par ALTree .

Nous nous sommes limités pour l'analyse aux 1000 premiers arbres équiparcimonieux obtenus par PAUP . Ces arbres enracinés sur la séquences consensus sont caractérisés par un CI de 0,21.

Test d'association Le test d'association a été réalisé sur trois arbres tirés aléatoirement parmi les 1000 arbres possibles. La p-value a été évaluée par 10 000 permutations. Elle est très légèrement supérieure au seuil de signification (p_{consH} comprise entre $0,06$ et $0,08$ pour les trois arbres). Comme cette valeur est très proche du seuil, nous avons tout de même recherché le ou les sites de susceptibilité potentiellement impliqués.

Localisation des sites de susceptibilité Pour l'attribution des états du caractère S, étant donné les faibles tailles d'échantillon pour chaque catégorie haplotypique, nous avons fixé le paramètre ϵ à 0,5. Nous avons choisi de ne pas déterminer l'état du caractère S pour les catégories haplotypique ne contenant qu'un seul haplotype (attribution de l'état " ?", indéterminé).

Nous avons testé 6 conditions d'analyse en fonction de l'enracinement (non enraciné, utilisation de l'haplotype consensus comme séquence ancestrale, utilisation de l'haplotype majoritaire comme séquence ancestrale) et du mode d'optimisation des caractères (acctran/deltran). Les résultats sont résumés dans le Tableau 4.6.

Ces résultats montrent que l'allèle 2 du SNP 3 semble être un allèle à risque pour la schizophrénie. Nous avons alors procédé comme dans l'analyse des données sur la maladie de Crohn. Nous avons éliminé les haplotypes portant l'allèle 2 du SNP3 de l'échantillon et nous avons ré-analysé le jeu de données. Nous avons obtenu 26 catégories haplotypiques. Les arbres de parcimonie ont alors un CI de 0,39. L'analyse avec ALTree ne nous a pas permis de

Non enraciné		Consensus		Majoritaire	
acctran	deltran	acctran	deltran	acctran	deltran
$3_{1->2}$ **(0,59)**	$3_{1->2}$ **(0,92)**	$3_{1->2}$ **(0,507)**	$3_{1->2}$ **(0,892)**	$8_{1->2}$ (1.967)	$2_{2->1}$ (0.49)
$1_{2->1}$ (0.199)		$1_{2->1}$ (0.152)		$3_{1->2}$ **(1.885)**	$1_{2->1}$ (0.409)
				$2_{2->1}$ (0.573)	
				$2_{1->2}$ (0.535)	

Le tableau présente, pour chaque jeu de paramètres, les sites et les sens de changements (en indice) correspondant aux V_i positifs. Les valeurs entre parenthèses correspondent au V_i divisé par le nombre d'arbres analysés (ici, 1000).

TAB. 4.6: Résultats de l'analyse des données sur la schizophrénie en utilisant un arbre reconstruit par parcimonie

mettre en évidence une association avec la maladie dans ce jeu de données ($p_{consH} = 0.29$). Cependant, il faut noter que l'allèle 2 du SNP 3 est très fréquent et que lorsque les haplotypes porteurs de l'allèle à risque sont éliminés, il ne reste plus que 32 malades et 55 témoins dans l'échantillon. Dans ces conditions, le test est très peu puissant et il serait de toute façon extrêmement difficile de mettre en évidence une association avec la maladie.

4.3.3.2 Reconstruction de l'arbre par maximum de vraisemblance

Nous avons utilisé le logiciel PHYML pour reconstruire l'arbre. Nous avons voulu conserver l'information contenue dans le microsatellite, mais, comme ce logiciel ne permet pas d'utiliser ce type de marqueur pour reconstruire l'arbre, nous l'avons recodé sur 5 caractères. Pour conserver le caractère ordonné du microsatellite (seul le passage à un état voisin est possible), nous avons choisi le codage détaillé sur la Figure 4.7.

État initial	Nouveau codage				
c	c_1	c_2	c_3	c_4	c_5
1	1	1	1	1	1
2	2	1	1	1	1
3	2	2	1	1	1
4	2	2	2	1	1
5	2	2	2	2	1
6	2	2	2	2	2

FIG. 4.7: Système de codage du microsatellite

Avec ce système de codage, le passage de l'état 1 à l'état 4 implique 3 pas, ce qui correspond bien à un codage "ordonné" tel que celui utilisé pour la reconstruction de l'arbre par la méthode de parcimonie.

Les probabilités de changement d'état des caractères sont calculées par le logiciel PAML .

Test d'association Le test d'association a été réalisé sur l'arbre le plus vraisemblable obtenu avec PHYML . La p-value corrigée a été calculée pour 10 000 permutations sur l'arbre enraciné

sur la séquence consensus formée des allèles les plus fréquents. Comme dans le cas où l'arbre est reconstruit par une méthode de parcimonie, nous avons trouvé une p-value légèrement supérieure au seuil de signification ($p_{consH} = 0,06$)

Localisation des sites de susceptibilité Comme pour l'analyse de l'arbre de parcimonie, nous avons choisi de fixer le paramètre ϵ à 0,5 et de ne pas fixer les états du caractère S pour les catégories haplotypiques représentées par un seul haplotype. L'analyse a été réalisée sur un arbre non enraciné. Seuls deux SNPs donnent un V_i supérieur à zéro. Il s'agit des SNP $3_{1->2}$ ($V_i = 1,92$) et $6_{2->1}$ ($V_i = 1,88$).

4.3.4 Conclusion

Les résultats sont moins probants que ceux obtenus sur le jeu de données concernant la maladie de Crohn, probablement parce que l'effet du gène sur la maladie est moins fort. L'association n'est pas vraiment détectée et selon les paramètres utilisés pour l'analyse, on n'identifie pas toujours les mêmes marqueurs. Cependant, le sens de changement d'état $1 \rightarrow 2$ du marqueur 3 (Taq1 A1/A2) est quasiment toujours parmi ceux pour lesquelles le V_i est le plus élevé. On peut donc penser que ce polymorphisme, et, plus particulièrement l'allèle A2, est associé à la schizophrénie. Il est intéressant de noter que cet allèle est celui qui a été précédemment identifié par Dubertret *et al.* (2004) comme étant associé à la schizophrénie. Son rôle avait alors été mis en évidence en utilisant un TDT allélique sur les mêmes données que celles que nous avons étudiées ainsi qu'en utilisant un test allélique cas/témoins comparant les fréquences alléliques chez les cas précédemment étudiés et chez un nouvel échantillon de témoins.

Conclusion

L'identification des facteurs génétiques de risque impliqués dans la susceptibilité aux maladies multifactorielles est l'enjeu majeur de la génétique épidémiologique. Les nouvelles méthodes développées dans ce but doivent permettre de traiter au mieux l'information d'un nombre de marqueurs sans cesse croissant. Les méthodes basées sur des phylogénies répondent à ce problème en regroupant les haplotypes en fonction de leur histoire évolutive. Au cours de cette thèse, nous avons développé une méthode haplotypique basée sur la reconstruction de la phylogénie des haplotypes et nous l'avons implémentée dans le logiciel `ALTree` .

Ces dernières années, plusieurs méthodes de détection de l'association basées sur la reconstruction d'arbres des haplotypes ont été développées. La méthode `ALTree` que nous avons proposé dans ce travail en fait partie. Le test d'association mis au point dans `ALTree` repose sur une analyse emboîtée des groupes sur l'arbre. Cette stratégie correspond à celle employée par Templeton *et al.* (1987); Seltman *et al.* (2003) et Durrant *et al.* (2004). Plus récemment d'autres stratégies ont été proposées, comme l'analyse de toutes les partitions de l'arbre (Templeton *et al.*, 2005), ou d'un seul niveau de regroupement (Tzeng, 2005). Les méthodes bassées sur des phylogénies diffèrent aussi par le choix de l'objet qu'elles analysent. Comme Durrant *et al.* (2004), nous avons choisi de travailler sur un arbre enraciné alors que Templeton *et al.* (1987); Seltman *et al.* (2003); Templeton *et al.* (2005) ont choisi de travailler sur un réseau. La méthode de Tzeng (2005) est, quant à elle, relativement atypique car les relations entre les haplotypes fréquents formant la base sont ignorées, et ne sont donc pas constituées en réseau. Templeton *et al.* (1987) justifient leur choix de travailler sur un réseau par le fait que cela représente mieux les relations entre les haplotypes au niveau intraspécifique qu'un arbre enraciné. Cependant, l'analyse emboîtée ne prend pas en compte les boucles dans le réseau, alors que ce sont elles qui constituent l'information supplémentaire par rapport à un arbre. Dans `ALTree` , nous avons choisi d'enraciner les arbres car la définition des groupes emboîtés est beaucoup plus simple dans ce cas. En effet, il existe des situations ambiguës que l'algorithme de définition des groupes mis au point par Templeton *et al.* (1987) ne peut résoudre, comme le montre la Figure 5.1. C'est pourquoi nous avons choisi de travailler sur des arbres enracinés, au moins pour la méthode de détection de l'association qui nécessite la définition de groupes emboîtés. De plus, la définition d'une racine permet de donner à l'arbre une dimension évolutive qui est absente sur les arbres non enracinés, quoi que laissent penser les terminologies employées par certains auteurs.

Enfin, on peut noter que toutes ces méthodes ont été développées dans le but d'identifier une association entre un gène candidat et une maladie. Seuls `CLADHC` (Durrant *et al.*, 2004) et `ALTree` (Bardel *et al.*, 2005a, 2006) proposent en plus une méthode pour localiser le site

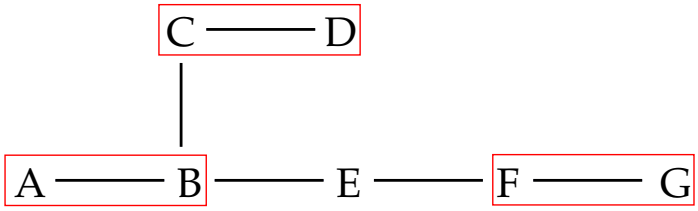


FIG. 5.1: Ambiguïtés de définition des groupes en utilisant l'algorithme de Templeton *et al.*

Cette figure présente un arbre non enraciné formé de 7 haplotypes différents. L'algorithme de Templeton permet de regrouper A et B, C et D, et F et G. L'haplotype E pose problème. Templeton et Sing (1993) proposent trois règles. En premier lieu, l'haplotype E doit être rassemblé avec le groupe dont l'effectif est le plus faible. Si plusieurs catégories possèdent le même effectif, l'haplotype doit être rajouté au groupe auquel il n'est pas relié par une branche ou un site de restriction apparaît où disparaît. Enfin, à défaut, le regroupement se fait aléatoirement.

de susceptibilité impliqué dans le déterminisme de la maladie.

Tout au long de notre travail, nous nous sommes intéressés à la puissance de ces méthodes pour détecter l'association. Nos simulations ont montré qu'`ALTree` et `CLADHC` n'étaient généralement pas plus puissants qu'un simple test d'association allélique réalisé à tous les locus. Ces résultats confirment les observations de Seltman *et al.* (2003) et de Tzeng (2005) montrant que l'utilisation de l'arbre ou du groupement des haplotypes n'entraîne pas un gain de puissance significatif par rapport à un test classique (TDT ou test haplotypique). Ces méthodes ont été développées avec l'idée sous-jacente que la majorité des haplotypes des malades formerait un nombre limité de groupes facilement identifiables sur l'arbre. En particulier, si une mutation unique est à l'origine de tous les haplotypes des malades, alors ces méthodes devraient être très puissantes. Cependant, dans le cas des maladies multifactorielles, nous sommes rarement dans cette situation car plusieurs facteurs génétiques ou environnementaux influent sur la maladie. Les malades ne forment donc généralement pas un groupe homogène. De plus, nous avons observé que les arbres reconstruits à partir des données — aussi bien pour CARD15 que pour DRD2 — contiennent énormément d'homoplasies. Une partie de l'homoplasie pourrait être imputable à des recombinaisons. Mais, même lorsque l'on sélectionne une région du gène CARD15 où l'on suppose qu'il y a peu de recombinaisons, le CI de l'arbre reste relativement faible (environ 0,3, ce qui signifie que chaque site mute en moyenne trois fois dans l'arbre). Ces observations sont en contradiction avec le modèle des sites infinis classiquement utilisé en génétique des populations selon lequel chaque site est supposé ne muter qu'une seule fois. Elles rejoignent celle de Templeton sur le gène de la lipoprotéine lipase (Templeton *et al.*, 2000) et montrent qu'il existe plusieurs régions sur le génome humain où un tel modèle ne s'applique pas. Les mutations de l'allèle à risque peuvent donc se produire sur des haplotypes différents. Les haplotypes des malades se répartissent alors dans plusieurs groupes, ce qui diminue la puissance des tests comparant les effectifs d'haplotypes portés par des malades et des témoins dans les différents groupes emboîtés.

Si la reconstruction de l'arbre n'apporte que peu d'amélioration à la détection de l'association, on constate en revanche qu'elle est très intéressante pour proposer des hypothèses

sur la localisation du locus de susceptibilité. Les simulations montrent que, pour des modèles impliquant plusieurs locus de susceptibilité, `ALTree` est bien plus efficace pour localiser les locus à risque qu'un simple test d'association allélique à chaque locus. L'application de la méthode développée dans `ALTree` à des données sur des maladies confirme son intérêt pour identifier des sites de susceptibilité : les trois variants du gène CARD15 impliqués dans le déterminisme de la maladie de Crohn ainsi que l'un des polymorphismes du gène DRD2 potentiellement impliqué dans la schizophrénie ont ainsi pu être confirmés.

Toutes les méthodes basées sur la reconstruction d'un arbre d'haplotypes supposent qu'il y a peu ou pas de recombinaisons dans la région étudiée car elles perturbent la reconstruction phylogénétique (Schierup et Hein, 2000). Lorsque cette condition n'est pas vérifiée, deux solutions ont été proposées. Elles nécessitent toutes deux que les évènements de recombinaison aient été détectés dans les données. La première solution consiste à reconstruire l'arbre sans s'occuper des recombinaisons, puis, à éliminer de l'arbre les haplotypes recombinants et tous leur descendants. On obtient finalement plusieurs arbres distincts : l'arbre initial dont on a enlevé des branches et les différentes branches formées des haplotypes recombinants et de leurs descendants. La seconde solution consiste à découper les haplotypes en fonction des points de recombinaison et à analyser séparément les différentes régions mises en évidence. C'est cette dernière méthode que nous avons mis en œuvre pour les données de la maladie de Crohn. Ces deux méthodes ont été décrites plus en détail par Templeton *et al.* (2000).

Cependant, jusqu'à présent, aucune étude n'a jamais quantifié l'impact réel de la recombinaison sur l'efficacité des méthodes basées sur la reconstruction d'arbres. Nous avons commencé à réaliser des simulations en utilisant un modèle de coalescence avec recombinaisons. Les premiers résultats montrent que, comme attendu, la présence de recombinaisons entraîne une baisse de puissance de la détection de l'association, mais cette baisse reste relativement limitée. D'autres analyses sont encore nécessaires pour confirmer ces premiers résultats ainsi que pour étudier l'influence de la recombinaison sur l'efficacité de la localisation du site de susceptibilité. Mais il semble qu'un taux de recombinaisons faible soit tout de même acceptable lors d'une analyse d'association basée sur une phylogénie. Avec ce type d'approche, quelques erreurs sur l'arbre ne portent pas à conséquence tant que les groupes d'individus malades peuvent être identifiés. Il est possible qu'en revanche, la recombinaison ait un effet plus important sur la localisation du site de susceptibilité.

Comme tous les tests haplotypiques, les méthodes basées sur l'utilisation d'un arbre sont confrontées au problème de la reconstruction des haplotypes. Hormis Seltman *et al.* (2003), tous les auteurs ont supposé, dans leurs études, que les haplotypes avaient été reconstruits sans ambiguïté. En pratique, des algorithmes de reconstruction haplotypique sont employés pour attribuer les haplotypes les plus probables à chaque individu. Or, aucune étude n'a jamais quantifié l'influence des méthodes de reconstruction des haplotypes sur les recherches d'association et sur la localisation des sites de susceptibilité basées sur des phylogénies. Nous pensons cependant que cette influence peut être relativement importante. En effet, lors de l'analyse des données sur la schizophrénie, on nous a tout d'abord fourni les haplotypes reconstruits "à la main", selon les critères personnels des auteurs de l'étude, les ambiguïtés de phase étant notées " ?". Nous avons comparé les résultats du test d'association réalisé avec `ALTree` sur les haplotypes ainsi obtenus et sur les haplotypes les plus vraisemblables reconstruits, pour chaque individu, avec le logiciel `FAMHAP` (Becker et Knapp, 2004a). Nous avons obtenu des p-values très différentes pour les deux tests ; le premier est très signifi-

catif alors que le second — que nous avons présenté dans ce travail — est à la limite de signification. Dans des travaux futurs, il sera donc important de quantifier l'impact de la reconstruction haplotypique sur la puissance de détection de l'association et sur l'efficacité de la localisation. Il faudra aussi adapter la méthode `ALTree` de manière à prendre en compte cette incertitude sur la reconstruction des haplotypes.

Parmi les méthodes utilisant des arbres, seule `ALTree` propose d'analyser des arbres reconstruits par maximum de vraisemblance. Cette méthode pourrait cependant être à l'origine de développements méthodologiques intéressants pour la localisation du site de susceptibilité. La méthode de maximum de vraisemblance nous donne les probabilités des différents états de caractères aux nœuds. Ces probabilités intègrent les ambiguïtés de reconstruction des états de caractères et évitent d'avoir à se poser la question du choix de la méthode d'optimisation des caractères (les résultats obtenus avec `acctran` ou `deltran` pouvant être très différents !). Pour l'instant, nous avons simplifié à l'extrême le problème en considérant que les états aux noeuds sont fixes : pour chaque caractère et à chaque nœud, l'état du caractère le plus probable est attribué. Nous avons vu au chapitre 3 que cette méthode n'est pas optimale par rapport à l'utilisation de la parcimonie. Une piste à explorer serait de tenir compte de la probabilité des états de caractères au noeud pour le calcul du V_i.

Dans le cas des maladies multifactorielles, la complexité des relations entre une maladie et les gènes qui en sont responsables peut être très importante. Un seul polymorphisme n'est pas nécessairement responsable de la maladie. Les outils méthodologiques développés pour la recherche de sites de susceptibilité isolés ne sont pas forcément bien adaptés à l'identification de plusieurs sites de susceptibilité en interaction, en particulier lorsque les effets marginaux des différents allèles sont faibles. Bien que plus efficace qu'un test locus par locus pour identifier des sites de susceptibilité en interaction, `ALTree` n'échappe pas à ce problème. De nouveaux développements méthodologiques sont donc nécessaires afin d'utiliser au mieux l'information phylogénétique en présence de sites en interaction. Une piste serait par exemple de considérer la phylogénie des haplotypes comme un outil pour sélectionner les marqueurs potentiellement intéressants. En effet, lorsque plusieurs sites en interaction sont responsables de la susceptibilité à une maladie, on s'attend à ce que les locus de susceptibilité aient tous muté sur les branches allant de la racine de l'arbre jusqu'au(x) groupe(s) contenant significativement plus d'haplotypes portés par des individus malades. On peut alors retenir l'ensemble des marqueurs ayant muté sur ces branches, puis les analyser par un test de combinaison (Jannot *et al.*, 2003; Jannot, 2004; Becker et Knapp, 2004b). L'association entre la maladie et toutes les combinaisons possibles des marqueurs étudiés est alors testée, ce qui permet de déterminer le SNP ou la combinaison de SNPs impliquées dans la susceptibilité à la maladie.

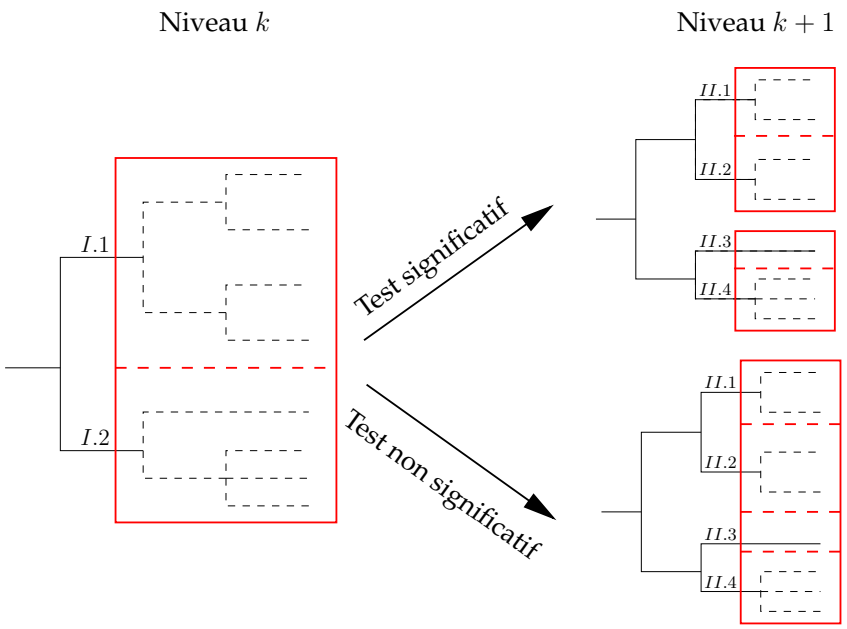
Finalement, au cours de cette thèse, nous avons souligné l'intérêt de l'utilisation des phylogénies dans les études de génétique épidémiologique pour permettre d'identifier des locus vraisemblablement impliqués dans la susceptibilité à la maladie. Cependant, pour des modèles génétiques où les effets du locus à risque sont faibles, le risque d'identifier un mauvais site peut parfois être assez élevé. Suite à ces études, il est donc nécessaire de réaliser des analyses fonctionnelles qui testeront si le où les polymorphismes détectés modifient l'expression ou le produit du gène d'une façon qui peut influencer la susceptibilité à la maladie.

Annexe A

Première méthode de localisation du site de susceptibilité

Principe de la méthode La première étape consiste à rechercher les niveaux de l'arbre auxquels on peut détecter une hétérogénéité de distribution des cas et des témoins. Pour cela, on réalise des tests d'homogénéité emboîtés selon le modèle ci-dessous (voir Figure A.1) :

- Si le test entre les i clades du niveau k est significatif, alors au niveau $k+1$, on réalisera i tests différents entre les descendants de chacun des i clades ;
- Si le test entre les i clades du niveau k n'est pas significatif, alors au niveau $k+1$, on réalisera un seul test entre tous les descendants des i clades.

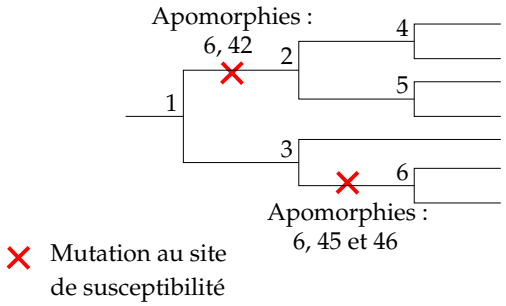


Au niveau k, un test d'homogénéité à 1 ddl est réalisé. S'il est significatif, deux tests à un ddl seront réalisés au niveau $k+1$. Sinon, un seul test à trois ddl sera réalisé au niveau $k+1$.

FIG. A.1: Réalisation des tests d'homogénéité

Ensuite, aux nœuds où une hétérogénéité de répartition des cas et des témoins est détectée, on réalise un test de proportion dans chacune des branches afin de rechercher l'origine de l'hétérogénéité et d'identifier d'éventuelles branches contenant un excès de cas par rapport au nœud qui les précède immédiatement dans l'arbre. Lorsqu'un clade contenant un excès de cas a été localisé, les synapomorphies définissant ce clade sont considérées comme des sites putatifs de susceptibilité à la maladie. Le parcours de l'arbre en aval se poursuit mais, cette fois, on recherchera des clades contenant un excès de témoins, ce qui pourrait correspondre à une éventuelle réversion de la mutation au site de susceptibilité.

Enfin, les sites putatifs de susceptibilité sont classés en fonction de leur probabilité d'être vraiment le site de susceptibilité. Pour cela, on calcule le rapport r entre le nombre de fois où ce site a été détecté et le nombre total de branches sur lesquelles on a détecté des sites de susceptibilité. Plus ce rapport est élevé, plus le site a de chance d'être impliqué dans le déterminisme de la maladie. Cette stratégie est illustrée sur la Figure A.2.



site 6 : $r=\frac{2}{2}$
site 42 : $r=\frac{1}{2}$
site 45 : $r=\frac{1}{2}$
site 46 : $r=\frac{1}{2}$
Le site 6 est donc le site putatif de susceptibilité le plus probable.

FIG. A.2: Identification des sites de susceptibilité les plus probables

Annexe B

Position des marqueurs et fréquence de leurs allèles dans les différentes régions du génome étudiées

B.1 Données de la base Seattles SNPs

	PLAU		IL13		TNF	
Marqueur	Position[a]	Freq a.m.[b]	Position[a]	Freq a.m.[b]	Position[a]	Freq a.m.[b]
1	82	0,30	293	0,19	282	0,17
2	211	0,28	693	0,19	352	0,06
3	835	0,02	2319	0,04	1009	0,06
4	3271	0,46	2346	0,02	1078	0,09
5	3272	0,46	3697	0,02	1440	0,06
6	3369	0,28	3727	0,26	1893	0,06
7	3855	0,11	3848	0,26	4040	0,06
8	3995	0,33	4329	0,24	4101	0,13
9	4625	0,46	4384	0,19	4366	0,06
10	4773	0,11	4553	0,22	4765	0,11
11	5634	0,52	6297	0,11	–	–
12	7358	0,43	6760	0,09	–	–
13	8061	0,43	–	–	–	–
14	8737	0,41	–	–	–	–
15	9170	0,43	–	–	–	–
16	9609	0,46	–	–	–	–

[a] Position en paires de bases
[b] Fréquence de l'allèle mineur

B.2 Données de la base Hapmap

	CHR20_1		CHR20_2	
Marqueur	Position[a]	Freq a.m.[b]	Position[a]	Freq a.m.[b]
1	0	0,45	0	0,44
2	36,5	0,39	73,2	0,20
3	44,9	0,42	119,8	0,11
4	88,1	0,49	185,3	0,50
5	124,1	0,10	193,9	0,27
6	156,1	0,21	230,8	0,48
7	206,5	0,07	287,3	0,50
8	288,9	0,34	300,6	0,38
9	295,2	0,43	306,9	0,34
10	331,1	0,23	311,2	0,09
11	379,4	0,14	351,2	0,18
12	384,4	0,47	441,9	0,35
13	461,4	0,33	–	–

Position en kilo paires de bases
Fréquence de l'allèle mineur

Bibliographie

Abecasis GR and WO Cookson. GOLD–graphical overview of linkage disequilibrium. *Bioinformatics*, 16(2):182–183, Mar 2000.

Adkins RM. Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genet*, 5(22), 2004.

Akey J, L Jin, and M Xiong. Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? *European J of Hum Genet*, 9:291–300, 2001.

Almasy L, JD Terwilliger, D Nielsen, TD Dyer, D Zaykin, and J Blangero. GAW12: Simulated genome scan , sequence, and family data for a common disease. *Genet epidemiol*, 21 (Suppl 1):S332–S338, 2001.

Bader JS. The relalative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics*, 2(1):11–24, 2001.

Bandelt HJ, P Forster, B Sykes, and MB Richards. Mitochondrial portraits of human populations using median networks. *Genetics*, 141:743–753, 1995.

Bardel C. Utilisation de phylogénies d'haplotypes pour la mise en évidence de facteurs génétiques de risque dans les maladies complexes, 2002. Mémoire de DEA de Biodiversité.

Bardel C, V Danjean, JP Hugot, P Darlu, and E Génin. On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genetics*, 6(24), 2005a.

Bardel C, P Darlu, and E Génin. Clustering of haplotypes based on phylogeny: how good a strategy for association testing? *Eur J Hum Genet*, 2005b. in press.

Bardel C, V Danjean, and E Génin. ALTree: Association detection and Localization of susceptibility sites using haplotype phylogenetic Trees. *Bioinformatics*, 2006. submitted.

Becker T and M Knapp. Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol*, 2004a.

Becker T and M Knapp. A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet*, 75(4):561–570, Oct 2004b. `http://www.uni-bonn.de/%7Eumt70e/becker.html`.

Bourgain C, E Génin, H Quesneville, and F Clerget-Darpoux. Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet*, 64:255–265, 2000.

Bourgain C, E Genin, P Holopainen, K Mustalahti, M Maki, J Partanen, and Clerget-Darpoux. Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *Am J Hum Genet*, 68(1):154–159, 2001.

Bourgain C, E Génin, and F Clerget-Darpoux. Comparison of family based haplotype methods using intragenic snps in candidate genes. *Eur J Hum Genet*, 10(5):313–319, 2002.

Brant SR and YY Shugart. Inflammatory bowel disease gene hunting by linkage analysis: rationale, methodology and present status of the field. *Inflamm bowel dis*, 10:300–311, 2004.

Burton PR, MD Tobin, and JL Hopper. Key concepts in genetic epidemiology. *The Lancet*, 366:941–951, 2005.

Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet*, 65(4):1170–1177, Oct 1999. `http://www-gene.cimr.cam.ac.uk/clayton/software/`.

Clement M, D Posada, and KA Crandall. Tcs: a computer program to estimate gene genealogies. *Mol Ecol*, 2000.

Clerget-Darpoux F, C Bonaïti-Pellié, and J Hochez. Effect of misspecifying genetic parameters in lod score analysis. *Biometrics*, 42:393–399, 1986.

Consortium TIH. The international hapmap project. *Nature*, 426:789–796, 2003. http://www.hapmap.org/.

Cooper DN and M Krawczak. *Human gene mutation*. New York: academic press, 1999.

Cordell HJ and DG Clayton. A unified stpwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type I diabetes. *Am J Hum Genet*, 70:124–141, 2002.

Crandall KA and AR Templeton. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics*, 134:959–69, 1993.

Culverhouse R, BK Suarez, J Lin, and T Reich. A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet*, 70:461–471, 2002.

Daly MJ, L Kruglyak, S Pratt, N Houstis, MP Reeve, A Kirby, and ES Lander. Genehunter, version 2.1, 2001a.

Daly MJ, JD Rioux, SF Schaffner, TJH TJ, and ES Lander. High-resolution haplotype structure in the human genome. *Nat Genet*, 29(2):229–232, 2001b.

Darlu P and E Génin. Cladistic analysis of haplotypes as an attempt to detect disease susceptibility. *Genet Epidemiol*, 21 (Suppl.1):S602–S607, 2001.

Darlu P and P Tassy. *Reconstruction phylogénétique, concepts et méthodes*. Masson, 1993.

Dempster AP, NM Laird, and DB Rubin. Maximum lakelihood from incomplete data via the em algorithm. *J R Stat Soc*, 39:1–38, 1977.

Dubertret C, P Gorwood, L Gouya, JC Deybach, and J Ades. Association and excess of transmission of a drd2 haplotype in a sample of french schizophrenic patients. *Schizophr Res*, 15(49):203–212, 2001.

Dubertret C, L Gouya, N Hanoun, JC Deybach, J Hadès, M Hamon, and P Gorwood. The 3' region of the drd2 gene is involved in genetic susceptibility to schizophrenia. *Schizophr Res*, 67:75–85, 2004.

Durrant C, KT Zondervan, LR Cardon, S Hunt, P Deloukas, and AP Morris. Linkage disequilibrium mapping via cladistic analysis of single-nucleotique polymorphism haplotypes. *Am J Hum Genet*, 75:35–43, 2004.

Falk CT and P Rubinstein. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet*, 1987.

Farris JS. Methods for computing Wagner trees. *Syst Zool*, 19:83–92, 1970.

Feingold J, M Fellous, and M Solignac. *Principes de génétique humaine*, chapter Liaison génétique. Hermann – éditeur de sciences et des arts, 1998.

Felsenstein J. *Inferring phylogenies*, chapter Chapter one: Parsimony methods, pages 1–18. Sinauer Associates, Inc., 2004a. ISBN: 0878931775.

Felsenstein J. *Inferring phylogenies*. Sinauer Associates, Inc., 2004b.

Felsenstein J. Phylip (phylogeny inference package) version 3.6. `http://evolution.genetics.washington.edu/phylip.html`, 2004c. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ferjerman L, N Bouzekri, X Wu, A Adeyemo, A Luke, X Zhu, R Ward, and RS Cooper. Association between evolutionary history of angiotensinogen haplotypes and plasma levels. *Hum Genet*, 115:310–318, 2004.

Friedlander Y, EM Berry, S Eisenberg, Y Stein, and E Leitersdorf. Plasma lipids and lipoproteins response to dietary challenge: analysis of four candidate genes. *Clin Genet*, 47:1–12, 1995.

Gabriel SB, SF Scaffner, H Nguyen, JM Moore, J Roy, B Blumenstiel, J Higgins, M DeFelice, A Lochner, M Faggart, SN Liu-Cordero, C Rotimi, A Adeyamo, R Cooper, R Ward, ES Lander, MJ Daly, and D Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

Ge Y, S Dudoit, and TP Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12:1–77, 2003.

Glatt SJ, SV Faraone, and MT Tsuang. Meta-analysis identifies an association between the dopamine d2 receptor gene and schizophrenia. *Mol Psychiatry*, 8(11):911–915, 2003.

Grassly NC and A Rambaut. Treevolve, version 1.32, 2000.

Guindon S and O Gascuel. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst biol*, 52:696–704, 2003.

Gusfield D. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In *6th international conference on computational biology, RECOM'02*, pages 166–168, 2002.

Hallman DM, S Visvikis, J Steinmetz, and E Boerwinkle. The effect of variation in the apoliprotein B gene on plasma lipid and apoliprotein B levels - I. A likelihood-based approach to cladistic analysis. *Ann Hum Genet*, 58:35–64, 1994.

Halperin E and E Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20:1842–1849, 2004.

Haviland MB, AM Kessling, J Davignon, and CF Sing. Cladistic analysis of the Apolipoprotein `ai-ciii-aiv` gene cluster using a healthy French Canadian sample. I. Haploid analysis. *Ann Hum Genet*, 59:211–231, 1995.

Hendy MD and D Penny. Branch and bound algorithms to determine minimal evolutionary tees. *Mathematical Biosciences*, 60:133–142, 1982.

Hudson RR. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18:333–338, 2002.

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

Hugot JP, P Laurent-Puig, C Gower-Rousseau, JM Olson, JC Lee, L Beaugerie, I Naom, JL Dupas, AV Gossum, M Orholm, C Bonaïti-Pellié, J Weissenbach, CG Mathew, JE Lennard-Jones, A Cortot, JF Colombel, and G Thomas. Mapping of a susceptibility locus for crohn's disease on chromosome 16. *Nature*, 379:821–823, 1996.

Jannot AS. *Détection de facteurs de risque génétiques dans des maladies multifactorielles*. PhD thesis, Université Paris XI, 2004.

Jannot AS, L Essioux, M Reese, and F Clerget-Darpoux. Improved use of SNP information to detect the role of genes. *Genet Epidemiol*, 25:158–67., 2003.

Jannot AS, L Essioux, and F Clerget-Darpoux. Association in multifactorial traits: how to deal with rare observations? *Hum Hered*, 58:73–81, 2004.

Judson R and JC Stephens. Notes from the snp vs. haplotype front. *Pharmacogenomics*, 2(1): 7–10, Feb 2001.

Jönsson EG, A Sillen, M Vares, B Ekholm, L Terenius, and GC Sedvall. Dopamine d2 receptor gene ser311cys variant and schizophrenia: association study and meta-analysis. *Am J Med Genet B Neuropsychiatr Genet.*, 15(119):28–34, 2003.

Kaplan N and R Morris. Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. *Genet Epidemiol*, 20(4):432–457, May 2001.

Katsov H, AM Bennet, P Kehoe, B Wiman, M Gatz, K Blennow, B Lenhard, NL Pedersen, U de Faire, and JA Prince. A cladistic model of ACE sequence variation with implications for myocardial infarction, alzheimer disease and obesity. *Hum mol genet*, 13:2647–2657, 2004.

Keavney B, CA McKenzie, JMC Connell, C Julier, PJ Ratcliffe, E Sobel, M Lathrop, and M Farrall. Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum Mol Genet*, 7(11):1745–1751, 1998.

Kerem B, JM Rommens, JA Buchanan, D Markiewicz, TK Cox, A Chakravarti, M Buchwald, and LC Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 8(245): 1073–1080, 1989.

Kestler LP, E Walker, and EM Vega. Dopamine receptors in the brains of schizophrenia patients: a meta-analysis of the findings. *Behav Pharmacol*, 12(5):355–371, 2001.

Kittles RA, JC Long, AW Bergen, M Eggert, M Virkkunen, M Linnoila, and D Goldman. Cladistic association analysis of Y chromosome effects on alcohol dependence and related personality traits. *Proc Natl Acad Sci*, 96:4204–4209, 1999.

Kluge AG and JS Farris. Quantitative phyletics and the evolution of anurans. *Syst Zool*, 18: 1–32, 1969.

Lander ES and NJ Schork. Genetic dissection of complex traits. *Science*, 265:2037–2048, 1994.

Lewontin RC and K Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14:450–472, 1960.

Liu JS, C Sabatti, J Teng, BJ Keat, and N Risch. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res*, 11(10):1716–1724, 2001.

Lobos EA and RD Todd. Cladistic analysis of disease association with tyroxine hydroxylase: Application to manic-depressive disease and alcoholism. *Am J Med Genet*, 74:289–295, 1997.

Lobos E and RD Todd. Association analysis in an evolutionary context: cladistic analysis of the DRD2 locus to test for association with alcoholism. *Am J Med Genet*, 81:411–419, 1998.

Long AD and CH Langley. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res*, 9(8):720–731, Aug 1999.

McGuffin P, MJ Owen, and AE Farmer. Genetic basis of schizophrenia. *The lancet*, 346:678–682, 1995.

McPeek MS and A Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale genetic mapping. *Am J Hum Genet*, 65:858–875, 1999.

Molitor J, P Marjoram, and D Thomas. Application of bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. *Genet Epidemiol*, 25(2):95–105, 2003a.

Molitor J, P Marjoram, and D Thomas. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet*, 73(6):1368–1384, 2003b.

Morris AP, JC Whittaker, and DJ Balding. Bayesian fine-scale mapping of disease loci, by hidden markov models. *Am J Hum Genet*, 67(1):155–169, 2000.

Morris AP, JC Whittaker, and DJ Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet*, 70(3):686–707, 2002.

Morris RW and NL Kaplan. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol*, 23(3):221–233, 2002.

Niu T. Algorithms for inferring haplotypes. *Genet Epidemiol*, 27, 2004.

Ogura Y, DK Bonen, N Inohara, DL Nicolae, FF Chen, R Ramos, H Britton, T Moran, R Karaliuskas, RH Duerr, JP Achkar, SR Brant, TM Bayless, BS Kirschner, SB Hanauer, and GN andJ H Cho. A frameshift mutation in nOD2 associated with susceptibility to crohn's disease. *Nature*, 411:603–606, 2001.

Patil N, AJ Berno, DA Hinds, WA Barrett, JM Doshi, CR Hacker, CR Kautzer, DH Lee, C Marjoribanks, DP McDonough, BT Nguyen, MC Norris, JB Sheehan, N Shen, D Stern, RP Stokowski, DJ Thomas, MO Trulson, KR Vyas, KA Frazer, SP Fodor, and DRC DR. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 23(294):1719–1723, 2001.

Payami H, S Joe, NR Farid, V Stenszky, SH Chan, PPB Yeo, JS Cheah, and G Thomson. Relative predispositional effects (RPEs) of marker alleles with disease: HLA-DR alleles and Graves disease. *Am J Hum Genet*, 45:541–546, 1989.

Posada D, TJ Maxwell, and AR Templeton. Treescan: a bioinformatic application to search for genotype/phenotype association using haplotype trees. *Bioinformatics*, 21:2130–2132, 2005.

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

Risch NJ. Searching for genetic determinants in the new millennium. *Nature*, 405:847–586, 2000.

Rodriguez S, TR Gaunt, SD O'Dell, XH Chen, D Gu, E Hawe, GJ Miller, SE Humphries, and IN Day. Haplotypic analyses of the igf2-ins-th gene cluster in relation to cardiovascular risk traits. *Hum Mol Genet*, 13(7):715–725, 2004.

Roeder K, SA Bacanu, V Sonpar, X Zhang, and B Devlin. Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol*, 28:207–219, 2005.

Saitou N and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.

Schaid DJ, SK McDonnell, L Wang, JM Cunningham, and SNT SN. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am J Hum Genet*, 71 (4):992–995, 2002.

Schierup MH and J Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156:879–891, 2000.

Schork NJ, D Fallin, and S Lanchbury. Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin Genet*, 58:250–264, 2000.

Schreiber S, P Rosenstiel, M Albrecht, J Hampe, and M Krawczak. Genetics of crohn disease, an archetypal inflammatory barrier disease. *Nat Rev Genet*, 6:376–388, May 2005.

Schulze TG, K Zhang, YS Chen, N Akula, F Sun, and FJ McMahon. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum Mol Genet*, 13(3):335–342, 2004.

Schwartz R, BV Halldorsson, V Bafna, AG Clark, and S Istrail. Robustness of inference of haplotype block structure. *J Comput Biol*, 10(1):13–19, 2003.

SeattleSNPs. Nhlbi program for genomic applications, uw-fhcrc, seattle, wa. (http://pga.gs.washington.edu) [accessed october 2004].

Seeman P. Dopamine receptors and the dopamine hypothesis of schizophrenia. *Synapse*, 1: 133–152, 1987.

Seltman H, K Roeder, and B Devlin. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet*, 68:1250–1263, 2001.

Seltman H, K Roeder, and B Devlin. Evolutionary-based association using haplotype data. *Genet Epidemiol*, 25:48–58, 2003.

Sokal RR and CD Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.

Sokal RR and PH Sneath. *Principles of numerical taxonomy*. Freeman, San Francisco, 1963.

Spielman RS, RE McGinnis, and WJ Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, 1993.

Stephens M and P Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73(5):1162–1169, 2003.

Stephens M, NJ Smith, and P Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–989, 2001a.

Stephens M, NJ Smith, and P Donnelly. Reply to zhang et al. *Am J Hum Genet*, 2001b.

Strauss J, CL Barr, CJ George, B Devlin, A Vetro, E Kiss, I Baji, N King, S Shaikh, H Lanktree, M Kovacs, and JL Kennedy. Brain-derived neurotrophic factor variants are associated with childhood-onset mood disorder: confirmation in a hungarian sample. *Mol Psychiatry*, 10 (9):861–867, 2005.

Swofford D and GJ Olsen. *Molecular systematics*, chapter Phylogeny reconstruction. Sinauer associates inc., (Sunderland, Massachusetts), 1990.

Swofford DL. PAUP phylogenetic analysis using parcimony. version 4.0b10. Sunderland, Massachusetts: Sinauer Associates, 2002. Available at http://paup.csit.fsu.edu/.

Templeton AR. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. analysis of case/control sampling designs: Alzheimer's disease and the Apolipoprotein E locus. *Genetics*, 140: 403–409, 1995.

Templeton AR and CF Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analysis with cladogram uncertainty and recombination. *Genetics*, 134:659–669, 1993.

Templeton AR, E Boerwinkle, and CF Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics*, 117:343–351, 1987.

Templeton AR, CF Sing, A Kessling, and S Humphries. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics*, 120:1145–1154, 1988.

Templeton AR, KA Crandall, and CF Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, 132:619–633, 1992.

Templeton AR, KM Weiss, DA Nickerson, E Boerwinkle, and CF Sing. Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics*, 156(3):1259–1275, 2000.

Templeton AR, T Maxwell, D Posada, JH Stengard, E Boerwinkle, and CF Sing. Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics*, 169:441–453, 2005.

Thomas DC, IB Borecki, G Thomson, K Weiss, L Almasy, J Blangero, D Nielsen, J Terwilliger, D Zaykin, and J MacCluer. Evolution of the simulated data problem. *Genet Epidemiol*, 21 (Suppl.1:S325–S331, 2001.

Thomas DC, DO Stram, D Conti, J Molitor, and P Marjoram. Bayesian spatial modeling of haplotype associations. *Hum Hered*, 56(1-3):32–40, 2003.

Tzeng JY. Evolutionary-based grouping of haplotypes in association analysis. *Genet Epidemiol*, 28:220–231, 2005.

Tzeng JY, B Devlin, L Wasserman, and K Roeder. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet*, 72:891–902, 2003.

Van der Meulen MA and GJ te Meerman. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genet Epidemiol*, 14(6):915–920, 1997.

Vermeire S, G Wild, K Kocher, J Cousineau, L Dufresne, A Bitton, D Langelier, P Pare, G Lapointe, A Cohen, MJ Daly, and J Rioux. CARD15 genetic variation in a Quebec population: prevalence, genotype-phenotype relationship, and haplotype structure. *Am J Hum Genet*, 71:74–83, 2002.

Wall JD and JK Pritchard. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am J Hum Genet*, (3):502, 515 2003.

Wang GQ, MD Pietro, K Roeder, CK Heng, CH Bunker, RF Hamman, and MI Kamboh. Cladistic analysis of human Apolipoprotein A4 polymorphisms in relation to quantitative plasma lipid risk factors of coronary heart disease. *Ann of Hum Genet*, 67:107–124, 2003.

Westfall PH and SS Young. *Resampling-based multiple testing*. John Wiley & Sons, New York, 1993.

Xu H, X Wu, MR Spitz, and S Shete. Comparison of haplotype inference methods using genotypic data from unrelated individuals. *Hum Hered*, 58(2):63–68, 2004.

Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13:555–556, 1997. http://abacus.gene.ucl.ac.uk/software/paml.html.

Yang Z, S Kumar, and M Nei. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–1650, 1995.

Yu K, J Xu, DC Rao, and M Province. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. *Ann Hum Genet*, 69:577–589, 2005.

Zaykin DV, PH Westfall, SS Young, MA Karnoub, MJ Wagner, and MG Ehm. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered*, 53:79–91, 2002.

Zhai W, MJ Todd, and R Nielsen. Is haplotype block identification useful for association mapping studies. *Genet Epidemiol*, 27:80–83, 2004.

Zhang K, P Calabrese, M Nordborg, and F Sun. Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet*, 71(6):1386–1394, Dec 2002.

Zhang S, AJ Pakstis, KK Kidd, and H Zhao. Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *Am J Hum Genet*, 69(4):906–914, 2001.

Zhu X, CA McKenzie, T Forrester, DA Nickerson, U Broeckel, H Schunkert, A Doering, H Jacob, RS Cooper, and MJ Rieder. Localization of a small genomic region associated with elevated ACE. *Am J Hum Genet*, 67:1144–1153, 2000.

Zhu X, RS Cooper, G Chen, A Luke, and R Eltson. Localization of the Q1 mutation by cladistic analysis. *Genet Epidemiol*, 21 (Suppl.1):S594–S599, 2001.

Zöllner S and JK Pritchard. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169(2):1071–1092, 2005.

Liste des publications

Revues internationales avec comité de lecture

C. Bardel, V. Danjean, and E. Génin. ALTree : Association detection and Localization of susceptibility sites using haplotype phylogenetic Trees. *Bioinformatics*, Nov. 2005. Submitted.

C. Bardel, V. Danjean, J.-P. Hugot, P. Darlu, and E. Génin. On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genetics*, 6(24), May 2005. URL `http://www.biomedcentral.com/1471-2156/6/24`.

C. Bardel, P. Darlu, and E. Génin. Clustering of haplotypes based on phylogeny : how good a strategy for association testing ? *European Journal of Human Genetics*, Nov. 2005. Accepted for publication.

Communications orales lors de conférences internationales avec comité de lecture

C. Bardel, V. Danjean, E. Génin, and P. Darlu. On the use of haplotype phylogenies to detect disease susceptibility loci, Apr. 2003. 31st European Mathematical Genetics Meeting, Cambridge (United Kingdom).

C. Bardel, V. Danjean, E. Génin, and P. Darlu. Efficiency of the cladistic approach to localise interacting susceptibility loci, Mar. 2004. 32nd European Mathematical Genetics Meeting, Kiel (Germany).

C. Bardel, P. Darlu, and E. Génin. Identification of disease susceptibility loci using a coevolution measure and haplotype phylogenies, Oct. 2005. 14th annual meeting of the International Genetic Epidemiology Society, Park City, USA. **Winner of the Williams Award of the International Genetic Epidemiology Society**.

P. Croiseau, F. Clerget-Darpoux, M.-C. Babron, C. Bardel, and E. Génin. Caution in the interpretation of association in presence of missing data, Apr. 2005. 33rd European Mathematical Genetics Meeting, Le Kremlin-Bicêtre, (France).

Posters lors de conférences internationales avec comité de lecture

C. Bardel, V. Danjean, P. Darlu, and E. Génin. Efficiency of the cladistic association analysis. Genetic Epidemiology 25 :236-279, abstract n°9, Nov. 2003. 12th annual meeting of the International Genetic Epidemiology Society, Redondo Beach (USA).

C. Bardel, V. Danjean, E. Génin, and P. Darlu. An application of phylogeny in genetic epidemiology : the detection of disease susceptibility loci, June 2005. Mathematics of Evolution and Phylogeny Conference, Paris (France).

C. Bardel, C. Dubertret, E. Génin, and P. Darlu. Detection of a variant in the DRD2 gene involved in schizophrenia using a phylogeny-based method, Oct. 2005. Annual Meeting of the American Society of Human Genetics, Salt Lake City (USA).

Posters lors de conférences nationales avec comité de lecture

C. Bardel, V. Danjean, P. Darlu, and E. Génin. Utilisation de phylogénies d'haplotypes pour la détection de facteurs génétiques de risque, Jan. 2004. 2ème Assises de Génétiques Humaine et Médicale, Angers (France).

Index

Les numéros en **gras** renvoient aux pages où les notions sont définies.

A

`acctran` ... voir *optimisation des caractères*/`acctran`
`ALTree` ... v–vii, 2, 61–63, 67, 68, 71, 75, 76, 91, 92, 94, 96–99, 101–104, 121, 122, 129, 136, 139–142
`ALTree-add-S` ... 67, 68
`ALTree-convert` ... 67, 68

B

Bonferroni (correction de) ... **18**, 42, 46, 48, 63, 65, 105

C

CI (indice de cohérence) ... **27**, 122, 129, 130, 136
`CLADHC` ... vi, vii, 2, 35, 44–46, 48, 58, 75, 76, 102, 104, 105, 139, 140
combinaison (test de) ... **13–15**, 142
correction de Bonferroni ... voir *Bonferroni (correction de)*

D

déséquilibre de liaison ... **5–6**, 8, 10–12, 36, 47, 48, 52, 93, 102, 103, 105, 122, 124, 129, 130, 134, 135
`deltran` ... voir *optimisation des caractères*/`deltran`
distance (méthode de) ... voir *méthodes de reconstruction phylogénétique*/*distance*

E

`EHAP` ... vii, 35, 41–44, 55, 57, 58

F

`FAMHAP` ... 8, 13, 68, 134, 141

G

`GENEHUNTER` ... 8, 122, 128
`GOLD` ... 129, 134
GRR (risque relatif génotypique) ... **9**, 47, 48, 94–96

H

homoplasie ... 7, **26**, 27, 53, 129, 130, 136, 140

I

indice de cohérence ... voir *CI*

M

méthodes de reconstruction phylogénétique
distance ... 25, **29–30**, 33, 34, 45, 59, 105
maximum de vraisemblance ... 25, **30–33**, 34, 62, 67, 74, 96, 98–99, 126–127, 131–133, 137–138, 142
parcimonie ... 25, **25–29**, 34, 37, 49, 53, 62, 67, 73–74, 96, 98–99, 122–124, 127, 129–131, 136–137, 142
maximum de vraisemblance (méthode de) ... voir *méthodes de reconstruction phylogénétique/maximum de vraisemblance*

N

NETWORK ... 53

O

optimisation des caractères
deltran ... **27**, 29, 74, 98–100, 102, 124, 136, 137, 142
acctran ... **27**, 29, 74, 91, 93, 95, 98–100, 102, 124, 136, 137, 142

P

pénétrance ... **8**, 12, 18, 47, 73, 75, 91–95, 97, 99, 101–103, 105, 127
PAML ... 33, 67, 74, 126, 131, 137
parcimonie (méthode de) ... voir *méthodes de reconstruction phylogénétique/parcimonie*
PAUP ... 26, 29, 33, 37, 53, 56, 58, 67, 68, 73, 97, 122, 123, 129–131, 136, 153
permutation
Ge *et al.* ... 14, **20**, 42, 63
simple ... **18–19**
Westfall et Young ... **19**, 50
PHASE ... 7, 68, 71–73
PHYLIP ... 29, 33, 37, 53, 56, 67, 68
PHYML ... 33, 34, 74, 97, 126, 131, 137

R

régression (méthodes de) ... **13**, 16, 41, 45, 46, 54
recombinaison ... 6, 9, 16–18, 37, 40, 42, 44, 45, 52, 55, 129, 140, 141

S

synapomorphie ... **26**, 27, 63

T

TCS ... 37, 49, 53, 56, 57
TDT
génotypique/allélique ... **12**, 14, 138
haplotypique ... 14, **15**, 41, 42, 140
test
test de combinaison ... voir *combinaison (test de)*
TRANSMIT ... 42

`TreeScan` 49, 50
`TREEVOLVE` 71, 75, 96, 102

Index des auteurs cités

Les numéros en **gras** correspondent aux articles inclus dans ce document.

A

Abecasis et Cookson (2000) 129, 134
Adkins (2004) 7
Akey *et al.* (2001) 16
Almasy *et al.* (2001) 122

B

Bader (2001) 16
Bandelt *et al.* (1995) 53, 57
Bardel *et al.* (2006) 68, **69**, 139
Bardel *et al.* (2005a) 76, **77**, 130, 139
Bardel *et al.* (2005b) 106, **107**
Bardel (2002) 63, 65
Becker et Knapp (2004a) 8, 134, 141
Becker et Knapp (2004b) . . . 13, 20, 68, 142
Bourgain *et al.* (2002) 10
Bourgain *et al.* (2000) 6, 10, 22
Bourgain *et al.* (2001) 10
Brant et Shugart (2004) 128
Burton *et al.* (2005) 4

C

Clayton (1999) 15, 42
Clement *et al.* (2000) 37, 49, 53, 57
Clerget-Darpoux *et al.* (1986) 9
Consortium (2003) 8, 105
Cooper et Krawczak (1999) 4
Cordell et Clayton (2002) 13
Crandall et Templeton (1993) . . . 38, 41, 53
Culverhouse *et al.* (2002) 16

D

Daly *et al.* (2001a) 122, 128
Daly *et al.* (2001b) 10
Darlu et Génin (2001) 53–55, 58
Darlu et Tassy (1993) 25, 30
Dempster *et al.* (1977) 7
Dubertret *et al.* (2001) 134
Dubertret *et al.* (2004) 134, 138
Durrant *et al.* (2004) . . . 2, 44–48, 54, 55, 58, 59, 104, 105, 139

F

Falk et Rubinstein (1987) 12
Farris (1970) 28
Feingold *et al.* (1998) 9
Felsenstein (2004a) 25
Felsenstein (2004b) 26
Felsenstein (2004c) 29, 37, 67
Ferjerman *et al.* (2004) 53–55, 57
Friedlander *et al.* (1995) 53

G

Gabriel *et al.* (2002) 10
Ge *et al.* (2003) 14, 20, 21, 42, 63
Glatt *et al.* (2003) 134
Grassly et Rambaut (2000) 75
Guindon et Gascuel (2003) 33, 34, 74
Gusfield (2002) 7

H

Hallman *et al.* (1994) 42, 53, 54, 56
Halperin et Eskin (2004) 7
Haviland *et al.* (1995) 53, 54, 56
Hendy et Penny (1982) 28
Hudson (2002) 52
Hugot *et al.* (2001) 53, 128, 131–133
Hugot *et al.* (1996) 128

J

Jönsson *et al.* (2003) 134
Jannot *et al.* (2004) 20, 22
Jannot *et al.* (2003) 13, 16, 20, 142
Jannot (2004) 142
Judson et Stephens (2001) 16

K

Kaplan et Morris (2001) 16
Katsov *et al.* (2004) 54, 57, 58
Keavney *et al.* (1998) 53–58, 127
Kerem *et al.* (1989) 17, 48
Kestler *et al.* (2001) 134
Kittles *et al.* (1999) 53, 54, 56, 58
Kluge et Farris (1969) 28

L

Lander et Schork (1994) 9
Lewontin et Kojima (1960) 5
Liu *et al.* (2001) 17, 18
Lobos et Todd (1997) 53, 54, 58
Lobos et Todd (1998) 53, 54, 58
Long et Langley (1999) 16

M

McGuffin *et al.* (1995) 133
McPeek et Strahs (1999) 17, 18
Molitor *et al.* (2003a) 17, 18
Molitor *et al.* (2003b) 17, 18
Morris et Kaplan (2002) 16
Morris *et al.* (2000) 17, 18
Morris *et al.* (2002) 17, 18

N

Niu (2004) 6, 7

O

Ogura *et al.* (2001) 128, 133

P

Patil *et al.* (2001) 10
Payami *et al.* (1989) 131
Posada *et al.* (2005) 49, 50

R

Rannala et Reeve (2001) 17
Risch (2000) 10
Rodriguez *et al.* (2004) 55, 57, 127
Roeder *et al.* (2005) 16

S

Saitou et Nei (1987) 30
Schaid *et al.* (2002) 8
Schierup et Hein (2000) 141
Schork *et al.* (2000) 5
Schreiber *et al.* (2005) 128
Schulze *et al.* (2004) 10, 11
Schwartz *et al.* (2003) 10
SeattleSNPs () 8, 52, 71, 73, 91, 105
Seeman (1987) 133, 134
Seltman *et al.* (2001) 41, 42, 44, 62, 104
Seltman *et al.* (2003) 41, 42, 54, 59, 139–141
Sokal et Michener (1958) 29, 30
Sokal et Sneath (1963) 30
Spielman *et al.* (1993) 12
Stephens et Donnelly (2003) 7, 68, 71
Stephens *et al.* (2001a) 7, 68, 71
Stephens *et al.* (2001b) 7
Strauss *et al.* (2005) 58
Swofford et Olsen (1990) 28
Swofford (2002) 29, 37, 67, 73

T

Templeton et Sing (1993) . . . 37, 38, 53, 55, 56, 140
Templeton *et al.* (1987) . . . 23, 36–38, 41, 53, 55, 56, 59, 62, 139, 140
Templeton *et al.* (1992) 37, 49, 53
Templeton *et al.* (2005) . . . 49, 50, 53, 54, 56, 57, 59, 139
Templeton *et al.* (1988) 37, 53, 56
Templeton *et al.* (2000) 140, 141
Templeton (1995) 37, 53, 58
Thomas *et al.* (2001) 122
Thomas *et al.* (2003) 17
Tzeng *et al.* (2003) 20, 23, 51
Tzeng (2005) 51, 52, 59, 139, 140

V

Van der Meulen et te Meerman (1997) . . . 10
Vermeire *et al.* (2002) 129, 133

W

Wall et Pritchard (2003) 11
Wang *et al.* (2003) 53, 54, 56

Westfall et Young (1993) 19, 50

X

Xu *et al.* (2004) 7

Y

Yang *et al.* (1995) 33
Yang (1997) 33, 67, 74
Yu *et al.* (2005) 22, 23

Z

Zöllner et Pritchard (2005) 17, 18
Zaykin *et al.* (2002) 16
Zhai *et al.* (2004) 11
Zhang *et al.* (2002) 16
Zhang *et al.* (2001) 7
Zhu *et al.* (2001) 53, 54, 56
Zhu *et al.* (2000) 53–56, 127

Table des figures

1.1 Le *Transmission Disequilibrium Test* . . . 12
1.2 Principe de la régression logistique ascendante pas à pas . . . 14
1.3 La permutation de Westfall et Young (1993) . . . 19
1.4 La permutation de Ge *et al.* (2003) . . . 21
1.5 Regroupement de catégories en utilisant un arbre de classification d'après Yu *et al.* (2005) . . . 22
1.6 Description d'un arbre phylogénétique non enraciné . . . 24
1.7 Différences entre un arbre enraciné par un ancêtre et un arbre non enraciné . 24
(a) Arbre non enraciné . . . 24
(b) Arbre enraciné . . . 24
1.8 Enracinement sur un groupe externe . . . 25
1.9 L'analyse par parcimonie . . . 26
1.10 Notion de convergence et de réversion . . . 27
(a) Convergence . . . 27
(b) Réversion . . . 27
1.11 Principe du *tree bisection and reconnection* (TBR) . . . 28
1.12 L'algorithme du neighbor-joining (NJ) . . . 31
1.13 Exemple d'arbre utilisé pour décrire le calcul de la vraisemblance . . . 32

2.1 L'analyse des groupes emboîtés de Templeton . . . 39
2.2 Localisation du site de susceptibilité en utilisant les haplotypes recombinants 40
2.3 Principe de la méthode de Seltman, Devlin et Roeder (EHAP) . . . 43
2.4 Découpage de l'arbre dans la méthode de Durrant *et al.* (CLADHC) . . . 46
2.5 Processus de simulation des haplotypes par Durrant *et al.* (2004) . . . 47
2.6 Principe de la méthode de Tree Scanning . . . 49
(a) Partition I . . . 49
(b) Partition II . . . 49
(c) Partition III . . . 49
(d) Partition IV . . . 49
(e) Partition II.1 . . . 49
(f) Partition II.2 . . . 49
(g) Partition II.3 . . . 49

3.1 Description de l'analyse emboîtée lors du test d'association . . . 64
(a) Niveau 1 . . . 64

(b) Niveau 2 . . . 64
(c) Niveau 3 . . . 64
3.2 Erreur de localisation du fait des petits échantillons . . . 65
3.3 Les différentes étapes d'une simulation . . . 72
3.4 Influence de la taille de l'échantillon sur le nombre de sites détectés avec le site de susceptibilité . . . 92
3.5 Influence de la fréquence de l'allèle de susceptibilité sur l'efficacité de la localisation . . . 93
3.6 Influence de la pénétrance de la maladie sur l'efficacité de la localisation . . . 95
(a) Gène *PLAU* . . . 95
(b) Gène *IL13* . . . 95
3.7 Influence de la méthode d'optimisation des caractères dans le cas où l'allèle de susceptibilité a une fréquence f faible ($f = 0,2$) . . . 102
3.8 Influence du déséquilibre de liaison maximum entre le site de susceptibilité et les autres marqueurs . . . 103
(a) Méthode locus par locus . . . 103
(b) Méthode `ALTree` . . . 103

4.1 Influence du choix de l'ancêtre sur la puissance du test d'association . . . 123
4.2 Exemple d'arbre illustrant les problèmes observés pour la localisation du site de susceptibilité pour les données GAW . . . 125
4.3 Organisation de la région contenant le gène CARD15 et localisation des 13 polymorphismes . . . 129
4.4 Analyse descriptive des données sur la maladie de Crohn . . . 130
(a) Fréquences alléliques des 13 SNPs chez les témoins et chez les malades 130
(b) Matrice de LD (mesure de déséquilibre : D') . . . 130
4.5 Organisation de la région DRD2/X-kinase et localisation des 8 polymorphismes étudiés. . . . 135
4.6 Analyse descriptive des données sur la schizophrénie . . . 135
(a) Fréquences alléliques des 7 SNPs . . . 135
(b) Matrice de LD (mesure de déséquilibre : D') . . . 135
4.7 Système de codage du microsatellite . . . 137
5.1 Ambiguïtés de définition des groupes en utilisant l'algorithme de Templeton *et al.* . . . 140

A.1 Réalisation des tests d'homogénéité . . . 143
A.2 Identification des sites de susceptibilité les plus probables . . . 144

Liste des tableaux

1.1 Exemple de table de contingence pour un test d'association génotypique . . . 12
1.2 Principe du TDT_{all} . . . 15
1.3 Localisation de la mutation impliquée dans la susceptibilité à la mucoviscidose 18

2.1 Revue des analyses phylogénétiques sur des données quantitatives . . . 57
2.2 Revue des analyses phylogénétiques sur des données qualitatives . . . 58

3.1 Caractéristiques des 3 gènes de la base Seattles SNPs . . . 72
3.2 Caractéristiques des deux régions du chromosome 20 . . . 73
3.3 Influence de la taille de l'échantillon sur l'efficacité de localisation . . . 91
3.4 Exemple de répartition des individus malades et témoins simulés dans les différentes catégories génotypiques pour différentes fréquences de l'allèle à risque . . . 94
3.5 Exemple de répartition des individus malades simulés dans les différentes catégories génotypiques pour différents modèles de maladie . . . 95
3.6 Influence de la méthode de reconstruction de l'arbre sur l'efficacité de localisation du site de susceptibilité . . . 97
3.7 Caractéristiques des vraisemblances des 142 arbres les plus parcimonieux (topologie $T5$) . . . 97
3.8 Influence du nombre d'arbres équiparcimonieux utilisés par `ALTree` . . . 98
3.9 Efficacité d'`ALTree` pour différentes méthodes d'optimisation des caractères 99
3.10 Influence de l'enracinement et du mode d'optimisation des caractères sur l'efficacité de la localisation avec `ALTree` . . . 101
3.11 Ecarts-types de la puissance de détection de l'association obtenus lors de l'analyse de 10 jeux de données simulés sous 4 modèles différents . . . 103

4.1 Localisation du site de susceptibilité sur les 208 arbres reconstruits par parcimonie (arbre non enraciné) . . . 124
4.2 Localisation du site de susceptibilité sur l'arbre de vraisemblance maximum (arbre non enraciné) . . . 126
4.3 Analyse globale du jeu de données Crohn sur des arbres reconstruits par parcimonie . . . 131
4.4 Résultats de l'analyse détaillée du jeu de données Crohn par parcimonie . . . 132
4.5 Analyse globale du jeu de données Crohn par maximum de vraisemblance . 132

4.6 Résultats de l'analyse des données sur la schizophrénie en utilisant un arbre reconstruit par parcimonie 137

Table des matières

Sommaire vii

Introduction 1

1 Introduction à la génétique épidémiologique et à la phylogénie 3
1.1 La génétique épidémiologique . . . 3
1.1.1 Les concepts de base . . . 4
1.1.1.1 Les polymorphismes . . . 4
1.1.1.2 La notion de pénétrance . . . 8
1.1.1.3 Les méthodes de recherche de facteurs de risque génétiques . 9
1.1.2 Les méthodes d'association en stratégie gène candidat . . . 11
1.1.2.1 Les différents tests d'association . . . 11
1.1.2.2 Les méthodes de localisation des sites de susceptibilité . . . 16
1.1.2.3 Tests multiples et catégories rares . . . 18
1.2 La phylogénie . . . 23
1.2.1 Les concepts de base . . . 23
1.2.1.1 Définition . . . 23
1.2.1.2 Les arbres phylogénétiques . . . 23
1.2.2 Les méthodes de reconstruction phylogénétique . . . 25
1.2.2.1 Les méthodes de parcimonie . . . 25
1.2.2.2 Les méthodes phénétiques . . . 29
1.2.2.3 La méthode de maximum de vraisemblance . . . 30
1.2.2.4 Les avantages et inconvénients de chaque méthode . . . 33

2 L'utilisation des phylogénies d'haplotype en génétique épidémiologique 35
2.1 L'analyse des clades emboîtés (Templeton *et al.*, 1987) . . . 37
2.1.1 Descriptif de la méthode . . . 37
2.1.1.1 Reconstruction phylogénétique . . . 37
2.1.1.2 Analyse de l'arbre . . . 38
2.1.2 Étude de puissance . . . 41
2.2 La méthode EHAP (Seltman *et al.*, 2003) . . . 41
2.2.1 Descriptif de la méthode . . . 41
2.2.1.1 Reconstruction phylogénétique . . . 41
2.2.1.2 Analyse de l'arbre . . . 41
2.2.2 Étude de puissance . . . 42

2.2.2.1 Résultats . . . 42
2.3 La méthode CLADHC (Durrant *et al.*, 2004) . . . 44
2.3.1 Descriptif de la méthode . . . 45
2.3.1.1 Reconstruction phylogénétique . . . 45
2.3.1.2 Analyse de l'arbre . . . 45
2.3.2 Étude de puissance . . . 46
2.3.2.1 Simulation des données . . . 47
2.3.2.2 Résultats . . . 47
2.3.3 Localisation du site de susceptibilité . . . 48
2.4 La méthode de Tree Scanning (Templeton *et al.*, 2005) . . . 49
2.4.1 Description de la méthode . . . 49
2.4.1.1 Reconstruction phylogénétique . . . 49
2.4.1.2 Analyse de l'arbre . . . 49
2.4.2 Étude de puissance . . . 50
2.5 La méthode de Tzeng (Tzeng, 2005) . . . 51
2.5.1 Descriptif de la méthode . . . 51
2.5.2 Étude de puissance . . . 52
2.5.2.1 Méthode de simulation . . . 52
2.5.2.2 Résultats . . . 52
2.5.2.3 Analyse de données . . . 53
2.6 Utilisation des méthodes phylogénétiques sur des données . . . 53
2.6.1 Reconstruction de l'arbre . . . 53
2.6.2 Type d'analyse statistique réalisée . . . 54
2.6.3 Détection de l'association . . . 54
2.6.4 Localisation des sites de susceptibilité . . . 55
2.7 Bilan sur les méthodes d'association et de localisation de sites de susceptibilité basées sur des phylogénies . . . 59

3 La méthode ALTree 61
3.1 Principe de la méthode ALTree . . . 62
3.1.1 Détection de l'association . . . 63
3.1.1.1 Principe du test . . . 63
3.1.1.2 Correction pour les tests multiples . . . 63
3.1.1.3 Gestion des arbres équiparcimonieux . . . 63
3.1.2 Localisation des sites de susceptibilité . . . 63
3.1.2.1 Principe de la méthode . . . 66
3.1.2.2 Gestion des arbres équiparcimonieux . . . 67
3.2 Implémentation de la méthode au sein du logiciel ALTree . . . 67
3.2.1 Résumé de l'article 1 : ALTree : Association detection and Localization of susceptibility sites using haplotype phylogenetic Trees . . . 67
3.2.2 Article 1 . . . 68
3.3 Étude de puissance . . . 71
3.3.1 Le processus de simulation . . . 71
3.3.1.1 Obtention d'un ensemble d'haplotypes . . . 71
3.3.1.2 Choix des sites de susceptibilité . . . 73
3.3.1.3 Tirage aléatoire des haplotypes et attribution d'un phénotype 73
3.3.1.4 Reconstruction phylogénétique . . . 73

3.3.1.5 Définition de la puissance de détection de l'association et de l'efficacité de localisation . . . 74
3.3.2 Résumé de l'article 2 : On the use of haplotype phylogeny to detect disease susceptibility loci . . . 75
3.3.3 Article 2 . . . 76
3.4 Les facteurs influençant la méthode . . . 91
3.4.1 La taille de l'échantillon . . . 91
3.4.2 Le modèle maladie . . . 92
3.4.2.1 La fréquence de l'allèle de susceptibilité . . . 92
3.4.2.2 La pénétrance de la maladie . . . 94
3.4.3 Les paramètres liés à la reconstruction phylogénétique . . . 96
3.4.3.1 La méthode de reconstruction de l'arbre . . . 96
3.4.3.2 Comparaison des méthodes de parcimonie et de vraisemblance pour l'optimisation des caractères . . . 98
3.4.3.3 L'enracinement et les méthodes d'optimisation des caractères 99
3.4.4 Variabilité liée aux jeux de données . . . 102
3.5 Étude de puissance d'une méthode phylogénétique alternative . . . 104
3.5.1 Nouvelle étude de puissance de la méthode de Durrant *et al.* . . . 104
3.5.1.1 Résumé de l'article 3 : Clustering of haplotypes based on phylogeny : how good a strategy for association testing ? . . . 104
3.5.2 Article 3 . . . 106

4 Application de la méthode à des données 121
4.1 Les données simulées pour GAW12 . . . 122
4.1.1 Description des données . . . 122
4.1.2 Analyse des données par `ALTree` . . . 122
4.1.2.1 Reconstruction de l'arbre par une méthode de parcimonie . . . 122
4.1.2.2 Reconstruction de l'arbre par une méthode de maximum de vraisemblance . . . 126
4.1.3 Conclusion . . . 126
4.2 La maladie de Crohn . . . 128
4.2.1 Le rôle du gène CARD15 . . . 128
4.2.2 Le jeu de données . . . 128
4.2.2.1 Obtention des données . . . 128
4.2.2.2 Analyse descriptive des données . . . 129
4.2.3 Analyse des données par `ALTree` . . . 129
4.2.3.1 Reconstruction de l'arbre par parcimonie . . . 130
4.2.3.2 Reconstruction de l'arbre par maximum de vraisemblance . . . 131
4.2.4 Conclusion . . . 133
4.3 La schizophrénie . . . 133
4.3.1 Le gène DRD2 : un bon gène candidat . . . 133
4.3.2 Le jeu de données . . . 134
4.3.2.1 Obtention des données . . . 134
4.3.2.2 Analyse descriptive des données . . . 134
4.3.3 Analyse des données par `ALTree` . . . 136
4.3.3.1 Reconstruction de l'arbre par parcimonie . . . 136
4.3.3.2 Reconstruction de l'arbre par maximum de vraisemblance . . . 137

4.3.4 Conclusion 138

Conclusion **139**

A Première méthode de localisation du site de susceptibilité **143**

B Position des marqueurs et fréquence de leurs allèles dans les différentes régions du génome étudiées **145**
B.1 Données de la base Seattles SNPs 145
B.2 Données de la base Hapmap 146

Bibliographie **147**

Liste des publications **157**

Index **159**

Index des auteurs cités **163**

Table des figures **167**

Liste des tableaux **169**

Table des matières **171**